

# Modelling Diffusion Processes in Social Networks and Visualizing Social Network Data

Juan Yao

A dissertation submitted in fulfilment of the requirements for the award of  
Master of Science



School of Computing  
DublinCityUniversity

Supervisor: Dr. Markus Helfert

September 2010

## Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Science is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: \_\_\_\_\_

(Candidate) ID No.: \_\_\_\_\_

Date: \_\_\_\_\_

## **Acknowledgements**

First of all, I would like to sincerely thank my supervisor, Dr. Markus Helfert, for many interesting discussions, patient corrections, constructive criticism and encouragement. Without his supervision this thesis would have never existed.

I would like to give special thanks to my colleague, Thoa Pham, Howard Duncan, Fakir Hossain for their help and patience.

I wish to thank EnterpriseIreland and Eircom for funding the project, School of Computing at the DublinCityUniversity for providing me a great working atmosphere.

In the end, I really want to thank my parents without whom I would never have gone this far. In the rest of my life, I can never finish paying back the love and care I owed to my parents.

## **Abstract**

There is a growing interest in exploring the role of social networks for understanding how individuals spread influence. In customer intelligence, social network analysis is fast emerging as an important discipline for predicting and influence consumer behaviour. Within online social networking, social media plays a prominent role in promotion, marketing and public relations. Hence, it is important to gain some deep insights of how information diffuses through social networks. In this context, this thesis presents an approach to approximating the diffusion process in social networks. It provides an alternative way of estimating the number of nodes reached by the initial target set in the diffusion process. Using this quantity as an influence measure, the proposed method can provide a way of identifying nontrivial nodes as influencers. The thesis conducts empirical analysis on a telecommunication phone user's network to assess social influence in the bundle adoption processes. This thesis also presents a tool for visualizing, exploring and manipulating social network data. The tool has rich support for dynamically manipulate and interactively exploring social network data. It also supports network visualization techniques such as satellite view, "lens effect" and visualizing social network data on a geographical map.

# Table of Contents

Declaration .....	I
Acknowledgements .....	II
Abstract .....	III
Table of Contents .....	IV
List of Tables.....	VIII
List of Figures .....	IX
Chapter 1 .....	1
Introduction .....	1
1.1 Motivations & Background.....	1
1.1.1 Motivation.....	1
1.1.2 Background & Research Questions .....	3
1.2 Contributions.....	5
1.3 Thesis Structure.....	6
Chapter 2 .....	7
Literature Review .....	7
2.1 Social Network Analysis.....	7
2.1.1 Development of Social Network Analysis.....	8
2.1.2 Social Network.....	9
2.1.3 Network Analysis.....	10
2.1.4 Network Analysis Goals & Methodology.....	14
2.1.5 Network Centrality Measures .....	16
2.1.5.1 Degree Centrality .....	17
2.1.5.2 Closeness Centrality.....	18
2.1.5.3 Betweenness Centrality .....	18
2.1.5.4 Eigenvector Centrality .....	20
2.1.6 Authority Ranking Measures .....	21
2.1.6.1 HITS Algorithm .....	21
2.1.6.2 PageRank .....	22
2.1.7 Clustering & Community Detection .....	23
2.1.7.1 Density, k-core & Cliques.....	25
2.1.7.2 Edge Betweenness Clustering.....	26

2.1.7.3 Voltage Cluster.....	27
2.2 Diffusion Processes in Social Networks .....	29
2.2.1 Modelling Diffusion Processes .....	30
2.2.1.1 The Dynamics of Adoption Process.....	31
2.2.1.2 Modelling Social Influence .....	32
2.2.1.3 Modelling Diffusion of Competing Technologies .....	33
2.2.2 Empirical Diffusion Studies.....	34
2.2.2.1 Characteristics of Social Influence .....	35
2.2.2.2 Patterns of Information Cascade .....	36
2.2.2.3 Role of Network Structure .....	37
2.2.3 Diffusion Application.....	38
2.3 Social Network Analysis Tools.....	38
2.4 Network Visualization .....	40
2.4.1 Graph Layout Algorithm.....	41
2.4.1.1 Criteria of Good Layout .....	41
2.4.1.2 Circular Layout .....	43
2.4.1.3 Geographic Layout.....	44
2.4.1.4 Force-directed Algorithms .....	44
2.4.1.5 Fruchterman-Reingold Layout.....	46
2.4.1.6 Kamada-Kawai Layout .....	48
2.4.1.7 Inverted Self Organising Map Layout.....	49
2.4.1.8 Tree Layout .....	49
Chapter 3 .....	52
Modelling Diffusion Processes in Social Networks .....	52
3.1 Motivations .....	52
3.2 Proposed Method .....	53
3.3 Experimental Validation .....	56
3.3.1 The Network Data .....	57
3.3.2 The experiments .....	57
3.4 Measuring Influence and Identifying Influencer.....	59
3.5 Discussion & Conclusion.....	61
Chapter 4 .....	63
Empirically Measuring Diffusion Processes.....	63
4.1 Introduction & Motivations .....	63

4.2 Data Description.....	64
4.3 Build Call Networks.....	65
4.4 Basic Network Characteristics .....	66
4.4.1 Degree Distribution.....	67
4.4.2 Connected Component Analysis.....	68
4.4.3 Persistence Analysis.....	69
4.5 Measuring the diffusion .....	70
4.5.1 Correlation Analysis .....	71
4.5.2 Dependence on Number of Friends Analysis.....	72
4.6 Summary and Discussion.....	75
Chapter 5 .....	77
Social Network Analysis & Visualization Tool.....	77
5.1 Introduction & Motivations .....	77
5.2 Overview of the Tool .....	79
5.2.1 Implementation Detail.....	79
5.2.2 Main Features.....	80
5.2.2.1 Network Data Manipulation.....	81
5.2.2.2 Visualization Features.....	83
5.3 Analysis Features .....	85
5.4 Analytical Visualization Features .....	87
5.4.1 Basic Analytical Visualization.....	88
5.4.2 Interaction and Navigation Techniques .....	89
5.4.2.1 Zooming .....	89
5.4.2.2 Interactive Filtering.....	90
5.4.2.3 Interactive Distortion .....	91
5.4.3 Visualize & Cluster Network Data Using Multidimensional Scaling. ....	92
5.4.4 Visualize Network Data on Geographical Map .....	96
Chapter 6 .....	99
Conclusion and Future Work .....	99
6.1 Summary & Contributions .....	99
6.1.2 Summary of Work.....	99
6.1.2 Contributions.....	100
6.2 Limitations .....	101
6.3 Future Work Direction .....	102

References ..... 104



## List of Tables

Table 1: Different social network representations. ....	10
Table 2: Sample Network Analysis & Visualization .....	12
Table 3: Diffusion modelling challenges and approaches. ....	30
Table 4: Summary of social network analysis tool types.....	39
Table 5: Summary network statistics for the sample customers' four month call networks. ....	67
Table 6: Size of the weakly connected components in the October 2008 call graph.	68
Table 7: Traditional SNA tool Pajek.....	78
Table 8: Process of visualizing network data using MDS. ....	94
Table 9: Screenshots of visualizing social network data on geographical maps. ....	98

## List of Figures

Figure 1: Example of which nodes have highest value of some centrality measures in a social network.....	17
Figure 2: An example of community structure in a social network.....	24
Figure 3: Cliques of Size 3, 4 and 5.....	26
Figure 4: Explicitly represent the step-by-step dynamics of adoption process.....	31
Figure 5: Probability of joining a community when $k$ friends are already members.....	35
Figure 6: Probability of buying DVD with incoming recommendations.....	35
Figure 7: A graph with random layout.....	42
Figure 8: A graph with circular layout.....	42
Figure 9: A graph drawing through a number of iterations of a force directed algorithm.....	45
Figure 10: An example of the spring embedder layout.....	47
Figure 11: An example of the tree layout.....	50
Figure 12: The plots show the time evolution of activated nodes count in the diffusion process with both simulation and calculation in the NS dataset.....	58
Figure 13: The plots show the time evolution of activated nodes count in the diffusion process with both simulation and calculation in the Hep-th dataset.....	58
Figure 14: The plot shows the distribution of influence value for all nodes in the NS dataset.....	60
Figure 15: The plot compares the influence value and node degree for all nodes in the NS dataset.....	60
Figure 16: The in-degree distribution of the sample customers' October 2008 call graph.....	68
Figure 17: The out-degree distribution of the sample customers' October 2008 call graph.....	68
Figure 18: Distribution of the sizes of the connected components in the October 2008 call graph.....	69
Figure 19: Distribution of nodes' overall persistence value.....	69
Figure 20: Distribution of the tie persistence value.....	70
Figure 21: Dependence on number of friends analysis.....	73
Figure 22: Distribution of $k$ value in set $U$ in join broadband event.....	74

Figure 23: Distribution of $k$ value in set $W$ in join broadband event.....	74
Figure 24: Distribution of the function $p(k)$ . .....	74
Figure 25: Main user interface of the prototypical tool. ....	80
Figure 26: User interface for setting node visualization properties. ....	83
Figure 27: Sample network with different node visualization properties.....	83
Figure 28: Screenshot of a network visualization with satellite view.....	84
Figure 29: Screenshot of a network visualization with lens effect. ....	85
Figure 30 : The edge betweenness community detection algorithm in my implementation.....	85
Figure 31: A network visualization with size of the node scaled to node degree.....	86
Figure 32: Differentiate groups by nodes colour and location.....	87
Figure 33: An example of the hyperbolic layout effect. ....	91
Figure 34: An example of the magnify view effect. ....	92
Figure 35: The MDS procedure output of a sample Network.....	95
Figure 36: User interface for setting the clustering options.....	95
Figure 37: A sample network with three clusters. ....	96

# Chapter 1

## Introduction

This chapter provides an introduction to the thesis. A brief theoretical background of the research is presented followed by a description of the motivation behind the work and the contributions provided by the thesis. Finally, an overview of the structure of the thesis is given.

### 1.1 Motivations & Background

In this section, firstly I present the motivations of the research. Then I introduce the background work and present my research questions.

#### 1.1.1 Motivation.

Social network analysis is not a brand new area. The first studies of social networks are attributed to J.L. Moreno in the 1930s who investigated how psychological wellbeing is related to the structural features of what he termed 'social configurations'. From the "six degrees of separation" theory (Milgram 1967) to Mark Granovetter's "the strength of weak ties" work on network structure (Granovetter 1973), and D.J. Watts' small world hypothesis (Watts and Strogatz 1998), social network analysis is a well-developed research area. However, many of these previous studies are based on small laboratory experiments with social networks limited to tens to hundreds of individuals. The extent of analysis has been limited due to the difficulty of obtaining data on very large-scale and completely representing social relationships. Recently, the emergence of mobile, email and online social networking have gradually transformed communication among people. Even as these new media have led to changes in our styles of communication, they have also remained governed by longstanding human social interaction principles. Social interaction through these platforms leaves extensive digital traces by its very nature. At unprecedented levels of scale and resolution we can now observe and quantify human communication patterns (Kleinberg 2008).

Recently, there is a growing interest in exploring the role of social networks for understanding how individuals spread influence. Customer intelligence is the process of gathering and analyzing information regarding customers; their details and activities, in order to build deeper and more effective customer relationships and improve strategic decision making. In customer intelligence, as more and more data revealing customer communication patterns is available in corporate data warehouse, social network analysis is fast emerging as an important discipline for predicting and influence consumer behaviour. Business analytics that leverages customer data through profiling, segmentation and predictive modelling is widely adopted by many customer-driven companies. Yet many companies still face low customer response rates to marketing initiatives, coupled with increasing customer churn. Research indicates that customers rely on each other's judgment and experience when making purchasing and loyalty decisions(Gabbott and Hogg 1994). Ignoring the network value of customers may lead to very suboptimal marketing strategies(Domingos and Richardson 2001).

In the context of customer intelligence, social network analysis is concerned with the analysis of customer behaviours in terms of how they interact with each other and who is influencing who on product purchase or service usage. Instead of focusing on individual customer characteristics, how the customers interact with each other is also important. It is argued that the value of analyzing customer networks is that the insight can help business practices from customer churn management to viral marketing campaigns(Doyle 2008). In Rob Cross's book – 'Driving Results through Social Networks: How Top Organizations leverage Networks for Performance and Growth', it filled with case studies for how leaders can understand and drive value through networks (Cross and Thomas, 2009).

On the other hand, within online social networking, social media plays a prominent role in promotion, marketing and public relations. Online social networking is booming and plays an ever important role in shaping user behaviours on the internet. Social network sites give users the ability to create individual profiles that foster interaction among people based on their interests, expertise or work activities. As a result of their massive popularity, these sites have been exploited as a platform for the dissemination of information, news and opinions(Cha, Mislove and

Gummadi2009). For instance, major movie studios place trailers for their movies on MySpace; US presidential candidates ran online political campaigns on YouTube; and individuals and amateur artists promote their songs, artwork, and blogs through these sites, all hoping to reach millions of online users. Hence it is important to gain some deep insights of how information disseminate across social networks.

### **1.1.2 Background & Research Questions**

Diffusion is a phenomenon in which new ideas or behaviours spread contagiously through social networks in the style of an epidemic. A rumour, a political message, or a link to an online video-these are all examples of information that can spread from person to person contagiously. It is a ubiquitous process in human social networks: A news story or an online video suddenly catches millions of attention; a new product or service gains sudden widespread popularity through word of mouth effect; the transmission of infective disease or computer virus. Many of these phenomena are very similar in principle - they tend to start with a few early adopters, and these early adopters may influence their friends, who may in turn influence their own friends and possibly lead to a cascade of influences(Kleinberg 2007).

An underlying premise many diffusion models build upon is assuming the existence of social influence – people are influenced by their neighbours in the social network when making adopting decision. Typically they formulate assumptions on how individuals respond to their friends' influence and further describe the way influence flows through the network. In this thesis I focus on models that explicitly represent the step-by-step dynamics of adoption. They assume the dynamic process unfolds in discrete time unit, with each node following certain probabilistic rule. For instance, an individual will adopt a new product or service when a certain threshold fraction of neighbours have already adopted (Granovetter 1978).

Among the many proposed models for diffusion process, two have garnered wide acceptance(Kempe, Kleinberg and Tardos 2003). In the Linear Threshold Model (LTM)(Granovetter 1978), each node is assigned a randomly chosen threshold, representing the fraction of neighbours required for it to adopt the new behaviour. A weight is assigned on each edge, indicating the extent of the influence. A node will adopt the behaviour if sum of the weights of its neighbours that have already adopted

the behaviour is greater than its threshold value. The other popular diffusion model is the Independent Cascade Model (ICM)(Goldenberg, Libai and Muller 2001)a probabilistic model in which a node catches the behaviour from its neighbours. In this model, when a node first becomes active it gives a single chance to activate its inactive neighbours with a probability - a parameter of the system.

While above models address the question of how influence spreads in a network, they are based on assumed rather than measured influence effects(Leskovec, Adamic and Huberman 2007). The wide variety of rules theoretical diffusion models posed on individuals' behaviour, even if plausible, are often lacking empirical support(Cointet and Roth 2007). Furthermore, most of the diffusion models take a single snapshot of the evolving network and then build upon this static network topology. Even though there are time steps in the simulation of the dynamic process, they have nothing to do the actual time of information spread. As such, it becomes unclear how accurately existing models render real-world diffusion phenomena. Cointet et al. (2007) suggest that future investigations of the diffusion mechanisms should begin with adequate empirical protocols; then propose adapted modelling frameworks.

Hence, it is necessary to propose new models that can capture the diffusion phenomenon more accurately. On the other hand as I stated earlier, collecting social network data has traditionally been hard work; many previous studies on social networks are based on small laboratory experiments. The emergence of social media has transformed the communication patterns among people. With the availability of such large-scale real network data, we can now measure the diffusion phenomenon at a more quantitative level. Indeed, it is important to obtain real world network data and conduct empirical analysis on such processes.

In this thesis, I attempt to extend the classic diffusion models to integrate some of the recent empirical findings on the diffusion phenomenon.I propose a new method to approximate the diffusion processes based on a Non-linear Dynamic System that describing disease spread in epidemiology.Meanwhile, I obtain data from a telecommunication operator and build call networks. I analyze the structure properties of the call graphs. Using broadband bundle adoption data I conduct empirical analysis to assess social influence in the diffusion processes.

Another area this thesis is interested in is visualizing social network data. Central to the development of social network analysis are two factors: measurement and visualization (Freeman 2000). Network visualization explores the patterns of relationships embedded in social networks by means of visual images. It is the process of creating a useful visual representation of a social network. It has played an important role in generating new insights in social network analysis. Good visualisation may reveal the hidden structure of the networks and amplifies human understanding, thus leading to new insights, new findings and possible prediction of the future. Moreover, the visualization of networks is important because it is a natural way to communicate connectivity and allows for fast pattern recognition by humans. It is useful for leveraging the perceptual abilities of humans to quickly explore and understand large amounts of data in parallel.

Within social network analysis research, there are many network analysis and visualization software both in commercial and open-source. So why create another tool? The reasons are twofold. First, many of the network analysis tools are designed for expert practitioners, have complex data handling, and inflexible graphing and visualization features that inhibit wider adoption. I believe it is necessary to come up with a simple tool that can allow users with little knowledge on social network analysis can perform analysis tasks. Second, many of the existing tools have extensive and complicated analysis and visualization features. I believe allowing users to interactively explore and manipulate network data should be a high priority. Motivated by these, my goal is to create an extendible network analysis tool that encourages interactive overview, discovery and exploration through “direct” data manipulation, graphing and visualization.

## **1.2 Contributions**

This thesis offers three principal contributions.

The first is the method for modelling the diffusion process based on an epidemic model that accurately models virus propagation in epidemiology. With proper parameters setting on a given network topology, the proposed method can numerically calculate each node's probability to get infected when a set of nodes has been initially activated. It provides an alternative way of estimating the number of



nodes reached by the initial target set in the diffusion process. Using this quantity as an influence measure, experiment results show that the proposed method can provide a way of identifying nontrivial nodes as influencer.

The second is the empirical study that assesses social influence in the telecommunication call networks with regard to the bundle adoption event. The analysis results show that the call networks are relatively stable during the examination period. Using correlation analysis the study confirms the existence of social influence in the call networks with regard to bundle adoption event. However, the characteristic of the influence is not same as it has been observed in other scenarios.

The third is the prototypical tool for social network data manipulation, analysis and visualization. The tool offers extensive features for users to interactively manipulate and explore social networks. It also provides distinctive visualization features such as allowing users to view the network data on a geographical map.

### **1.3 Thesis Structure**

The remainder of this thesis is structured as follows:

In chapter 2 various literatures related to my work on the diffusion processes in social networks and network analysis and visualization tool have been reviewed.

In chapter 3 I present my work on modelling the diffusion process in social networks.

In chapter 4 I present my work on empirically assessing the social influence and measuring diffusion processes in the telecommunication call networks.

In chapter 5 I introduce the prototypical tool for social network data manipulation, analysis and visualization.

Conclusions and suggestions for future work are given in Chapter 6.

## Chapter 2

### Literature Review

In this chapter, following my research motivations, I focus on literature related to the diffusion processes in social networks and network visualization. Since most of my work is conducted in the context of social network analysis, I start with an overview of social network analysis research. Furthermore, as I try to develop a network analysis tool, in section 2.1, I first investigate the methodology and goals of conducting network analysis then I provide theoretical background to the analysis techniques featured in my tool.

One of the major research areas this thesis is interested in is the diffusion processes within social networks. As mentioned earlier, diffusion processes within social networks has known an increased interest in recent years. An extensive amount of theoretical and empirical literature has been devoted to this phenomenon and the mechanisms behind it. There are many options and approaches. In section 2.2, I introduce various state of the art research issues and provide theoretical background for my work on modelling diffusion process and empirical diffusion studies. Specifically, I summarize some of the major challenges faced when modelling diffusion process and common approaches to address these challenges. I categorize recent empirical studies on diffusion process and summarize their findings.

After that, I present literature review related to my work on network analysis and visualization tool development. In section 2.3, I provide an overview of some of the major social network analysis tools, summarizing their advantages and disadvantages. In section 2.4, I introduce various graph layout algorithms that used to position the nodes and edges in the network visualization.

#### 2.1 Social Network Analysis

As most of my work is conducted in the context of social network analysis, I start with providing an overview of this area. Firstly, I examine what social networks are in section 2.1.1. I then introduce common notation and fundamental concepts of

general network analysis in section 2.1.2. In the literature, numerous measures have been defined to statistically assess every aspect of the social networks. However, there is no systematic way to interpret networks. In section 2.1.3, I examine the methodology and goals of performing network analysis. From this study, I have identified the two major objectives of network analysis is to uncover notable nodes and cohesive subgraphs. In section 2.1.4 and 2.1.5, I introduce various centrality and authority ranking measures to discover “important” nodes. In section 2.1.6, I present various concepts related to identify cohesive subgraphs and algorithms that detect communities in social networks.

### **2.1.1 Development of Social Network Analysis**

Social network analysis has emerged as a key technique in modern sociology. People have used the idea of “social network” loosely for over a century to connote complex sets of relationships between members of social systems that at all scale, from interpersonal to international. In social science, the structural approach that is based on the study of interaction among social actors is called social network analysis.

A summary of the development of social networks and social network analysis has been written by Linton Freeman (Freeman 2004). In his book, Freeman divides the history of the social network paradigm into four distinct periods. The first, which lasted from the mid-nineteenth century to the 1920s, is the origins of the social network ideas and practices. The second, which started in the late 1920s and run through the 1930s, is the birth of social network analysis. A major force behind the full-fledged emergence of network analysis was Jacob Levy Moreno. He pioneered the systematic recording and analysis of social interaction in small groups. The period stretching from the 1940s up to the 1960s Freeman described as a kind of “dark ages” for social network analysis. This was a time when social network analysis remained fragmented and localized. It was only in the 1970s, during a period Freeman describes as “the Renaissance at Harvard”, that social network analysis finally became a generalized paradigm for research. One group was centered around Harrison White and his student at the Harvard University were most responsible for this turn of event.

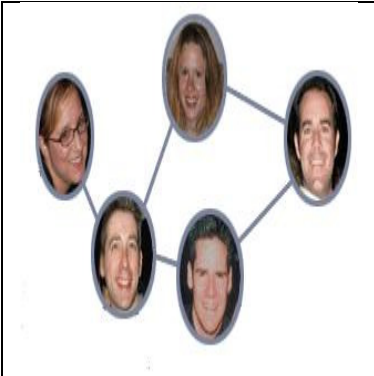
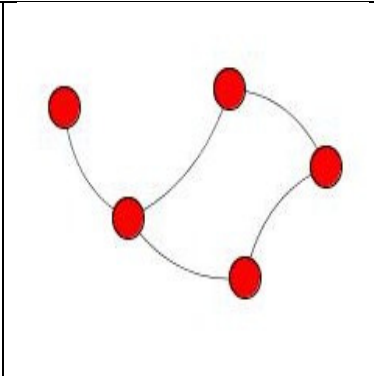
## 2.1.2 Social Network

Social network is a social structure containing various nodes (usually organizations and individuals) which are connected by one or more kinds of relationships. It indicates the ways in which they are connected through various social familiarities ranging from casual acquaintance to close familiar bonds (Hanneman and Riddle 2005). Email traffic, disease transmission and criminal activity can all be modelled as social networks. The nodes in the social network are called actors, and the links between each pair of actors are called relationships. A social network can have several different kinds of relationships and even multiple relationships simultaneously. Relations can be undirected or directed. For undirected relations, the link between two actors does not have a to or from consideration. For directional relations the path from node  $i$  to node  $j$  is different from the path from node  $j$  to node  $i$ . Relations may be weighted or unweighted, and weights, if present, may be interpreted as increasing or decreasing the tie between the two entities.

Social networks can be represented in a graph form. A graph usually denoted  $G(V, E)$  which consists of set of vertices  $V$  together with a set of edges  $E$ . In this context, vertices denote actors in the social network and edges denote the relationships between the actors. In this thesis, we use the term actor, node, vertex can interchangeably, and also the term edge and link. Based on the graph nature of social networks, the analysis of social network is highly dependent on graph theory. The main concepts about graph theory are assigned with walks, paths and distances. A sequence of adjacent vertices  $v_0, v_1, \dots, v_n$  is known as a walk. A walk in which no vertex occurs more than once is known as a path. A walk between two vertices whose length is as short as any other walk connecting the same pair of vertices is called a geodesic. The distance between two vertices is defined as the length of a geodesic that connects them. Traditionally there is no social network theory and the graph concepts can be employed in distinct social theories, requesting just some additional empirical and complementary information. The relations and individuals identification are mainly tasked to be performed in order to establish a real social network analysis.

Besides graph, a matrix is another common form of representing social network, where rows and columns as actors and the elements represent the ties between the actors. The simplest and most common matrix is binary. That is, if a tie is present, the cell value is one; if there is no tie, the cell value is zero. In a directed graph, by convention the sender of a tie is the row and the target of the tie is the column. One of the advantages of using a matrix to represent social relations is that it can easily allow the application of mathematical and computer tools to summarize and find patterns. In the following sections, we can find many matrix-based graph algorithms. Below in Table 1 I use a simple example to illustrate different network representations:

Table 1: Different social network representations.

		$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$
--	---	---

### 2.1.3 Network Analysis

Social network analysis (SNA) is the study of relations among a set of actors. The purpose of social network analysis is to identify important actors, crucial links, subgroups, roles, network characteristics, and so on, to answer substantive questions about structures. It has emerged as a powerful method for understanding the importance of relationships in networks. Freeman (2004) suggests that social network analysts seek to uncover two types of patterns in networks: one that shows subsets of nodes organized into cohesive social groups and others that uncover subsets of nodes having equivalent social positions or roles. Wasserman and Faust (2007) argue that SNA is based on assumption of the importance of relationships among interacting units. Network analysts focus on relationships instead of just the individual elements; how the elements are put together is just as important as the elements themselves. Prior to this perspective, social research focused largely on individual

attributes and neglected the social part of behaviour (how individual interact and the influence they have on each other.)(Freeman 2004)

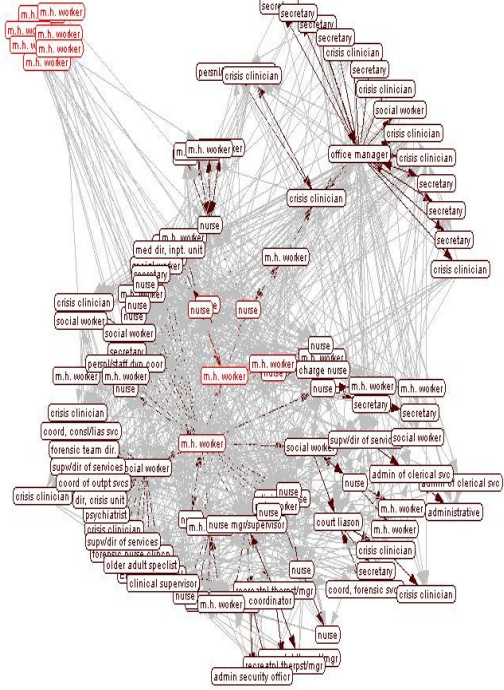
The power of social network theory stems from its difference from traditional social scientific studies is that it focuses on the value of the network structure rather than the characteristics of the individual. For instance, the network structure of an organization will affect its ability to access new ideas, recruit new individuals, and achieve sustainability. This approach has turned out to be useful for explaining many real-world phenomena.

The origin of contemporary social network analysis can be traced back to the work of Stanley Milgram (Milgram 1967). In his research, Milgram found that most people were connected by six acquaintances. This research led to the famous phrase “six degrees of separation”, which is still widely used in popular culture. Another important step in the development of social network analysis was the work of Mark Granovetter on network structures (Granovetter 1973). It is argued that weak social ties (distant and infrequent relationships) are more efficient at sharing knowledge as they provide access to novel information from otherwise disconnected parties. An example is people so often get job opportunities from their weak ties. D.J. Watts’ small world hypothesis builds upon both Milgram’s “six degrees of separation” concept and Granovetter’s “weak ties” argument by stating that most networks in the natural and man-made world are highly clustered yet far-reaching (Watts and Strogatz 1998). In these networks most nodes are not neighbours of one another, but most nodes can be reached from every other by a small number of hops or steps.

Understanding networks is an inherently difficult process. Within graph theory and social network analysis, various measures have been defined to statistically assess every aspect of social networks, such as size, density, degree, distance, diameter, centrality and etc. Using these techniques and metrics network analysts can find patterns in the structure, witness the flow of resources through a network, and learn how individuals are influenced by their surroundings. Social network analysis: Methods and Applications, by Wasserman and Faust, is perhaps the most widely used reference book for structural analysts (Wasserman and Faust 2007). The book presents a review of network analysis methods and an overview of the field. Below

in Table 2 is a sample analysis on a network generated by the Network Workbench Tool (Team 2006):

Table 2: Sample Network Analysis & Visualization

	<p>This graph claims to be directed.  Nodes: 113; Isolated nodes: 0  Node attributes present: label, area  Edges: 861  No self loops were discovered.  No parallel edges were discovered.  Edge attributes:      Did not detect any nonnumeric attribute.      This network seems to be valued.  Average total degree: 15.238938053097344  Average in degree: 7.619469026548669  Average out degree: 7.619469026548671  This graph is weakly connected.  There are 1 weakly connected components. (0 isolates)  The largest connected component consists of 113 nodes.  This graph is not strongly connected.  There are 13 strongly connected components.  The largest strongly connected component consists of 101 nodes.  Density (disregarding weights): 0.06803  Additional Densities by Numeric Attribute  densities (weighted against standard max)  weight: 0.06803  densities (weighted against observed max)  weight: 0.06803</p>
--	---

As seen from Table 2, from the network visualization on the left hand side, we can gain an initial impression on the structure of the network. From the network statistics on the right hand side, we can gain a clearer picture of the network structure: It consists of 113 nodes and 861 edges and it is directed. It is a weakly connected network with no isolated nodes. There are 13 strongly connected components and so on.

According to Wagner (2003), there are three main levels of interest: the element, group, and network level. On the element level, one is interested in properties of single actors, links, or incidences. Examples for this type of analysis are bottleneck identification and structural ranking of network items. On the group level, one is interested in classifying the elements of a network and properties of subnetworks. Examples are actor equivalence classes and cluster identification. Finally, on the network level, one is interested in properties of the overall network such as connectivity or balance. Here I briefly summarize the definitions of some of the common measures used to describe a social network or graph (Wasserman and Faust 2007):

**Degree:** the count of the number of ties to other actors in the network. Or in graph theory, the degree of a vertex of a graph is the number of edges incident to the vertex, with loops counted twice

**Loop:** A loop is an edge that connects a vertex to itself.

**Isolated node:** A vertex of zero degree, with no edges connected to it.

**Path:** is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence.

**Diameter:** is the greatest distance between any pair of vertices. To find the diameter of a graph, first find the shortest path between each pair of vertices. The greatest length of any of these paths is the diameter of the graph.

**Connected component:** is a subgraph in which any two vertices are connected to each other by paths. A directed graph is called strongly connected if there is a path from each vertex in the graph to every other vertex. In particular, this means paths in each direction; a path from a to b and also a path from b to a.

**Density:** is the number of edges in a network, expressed as a proportion of the maximum possible number of edge. Dense network means a network with relatively more edges and a complete network is a network with maximum density.

**Cluster coefficient:** A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'.

**K-core:** A k-core is a maximal subgraph in which each vertex has at least degree k within the subgraph.

**Clique:** a clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by edge.

**Closeness:** is the inverse of the sum of the shortest distances between each individual and every other person in the network. It measures the degree an individual is near all other individuals in a network (directly or indirectly).



**Betweenness:** vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not. It measures the extent to which a node lies between other nodes in the network.

**Centralization:** The difference between the number of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the number of links each node possesses.

**Structural equivalence:** Refers to the extent to which nodes have a common set of linkages to other nodes in the system. The nodes don't need to have any ties to each other to be structurally equivalent.

**Structural hole:** Static holes that can be strategically filled by connecting one or more links to link together other points. Linked to ideas of social capital: if you link to two people who are not linked you can control their communication.

#### **2.1.4 Network Analysis Goals & Methodology**

Numerous measures have been proposed by structural analysts to statistically assess social networks (Wasserman and Faust 2007). However, there is no systematic way to interpret networks, as measures can have different meaning in different networks. For instance, even though there are many importance rankings, clustering algorithms and statistical techniques for assessing social networks, there is no well-defined methodology for performing these operations (Perer and Shneiderman 2008b). In this section, I attempt to investigate the general goals and methodology for network analysis. During the design of my tool, I try to follow these guidelines so that users can assess the social networks in question more efficiently.

The purpose of social network analysis is to identify important actors, crucial links, subgroups, roles, network characteristics, and so on, to answer substantive questions about structures (Wagner2003). Some research questions focus on the structure of the whole graph or large sub-graphs, other questions focus on identifying individual nodes that are of particular interest. Some analysts will want to analyze the whole graph aggregated over its entire lifetime; others will want to slice the network into units of time to explore the progression of the network's development. According to

Perer and Shneiderman the 7-step methodology for social network analysis is (Perer and Shneiderman 2008b):

1. Overall network metrics, e.g., number of nodes, number of edges, density, diameter
2. Node rankings, e.g., degree, betweenness, closeness centrality
3. Edge rankings, e.g., weight, betweenness centrality
4. Node rankings in pairs, e.g., degree vs. betweenness, plotted on a scatter gram)
5. Edge rankings in pairs
6. Cohesive subgroups, e.g., finding communities
7. Multiplexity, e.g., analyzing comparisons between different edge types, such as friends vs. enemies.

This is not the only systematic method for social network analysis. At its most basic level, the initial step is to grasp basic insights into its overall network structure both statistically and visually. This may involve the overall network metrics such as density, diameter and number of components and are presented alongside a force-directed layout of the network. The visualization gives users a sense of the structure, clusters and depth of a network, while the statistics provide a way to both confirm and quantify the visual findings. Meanwhile from an overview, the network analysts can begin to seek sub-elements of the network that are of particular interest. This may involve finding sub-groups or cliques and measuring their cohesion in terms of the density of their internal connections.

Next step may include to gaining a deeper understanding of the individual elements of the network. Some nodes are notable, for example, because they have an extreme degree of connection to other nodes. For network analysts, identifying notable nodes in the network are of great importance and application. This may involve using statistical importance metrics common in social network analysis to measure the nodes and edges. For instance, an analyst can rank the nodes by degree (the most connected nodes), betweenness (the gatekeepers), closeness (well-positioned nodes to receive information) or other metrics. Some nodes play critical roles as bridges, sinks, or sources within the network. With an overview of a network in place,

network analysts may seek to find the gaps or holes in the network that could indicate missing data, a hidden actor, or a strategic gap that should be filled.

In the development of the network analysis module of my tool, I try to follow these principles. In the following sections, I investigate some of the measures in detail. In particular, I am interested in network centrality measures that rank network elements and identify “important” nodes; concepts and algorithms that uncover cohesive subgroups.

### **2.1.5 Network Centrality Measures**

Centrality is a group of measurements measuring how central a node is in a graph. In social networks, centrality measures are introduced as importance measures to quantify the role played by an actor in the complex interaction and communication occurring in the network. A central or prominent actor is defined as “those that are extensively involved in relationships with other actors. This involvement makes them more visible to the others.” (Wasserman and Faust 2007)

Being central in a graph can have many different meanings and therefore there are several different kinds of centrality measures. Three measures of centrality that are widely used in network analysis are degree centrality, betweenness centrality, closeness centrality. Although many other measures of centrality have been proposed since, these three continue to dominate. Here in Figure 1 I use a sample network with different highest centrality nodes pointed out to illustrate the idea. In the example, node 8 has highest betweenness centrality score; node 10 has highest closeness centrality score; node 13 has highest degree centrality score.

In the following section, I will introduce some of these most well-known centrality measures, as well as a number of link topological ranking measures. Some of these measures exhibit high computational complexity, while others are simple to calculate.

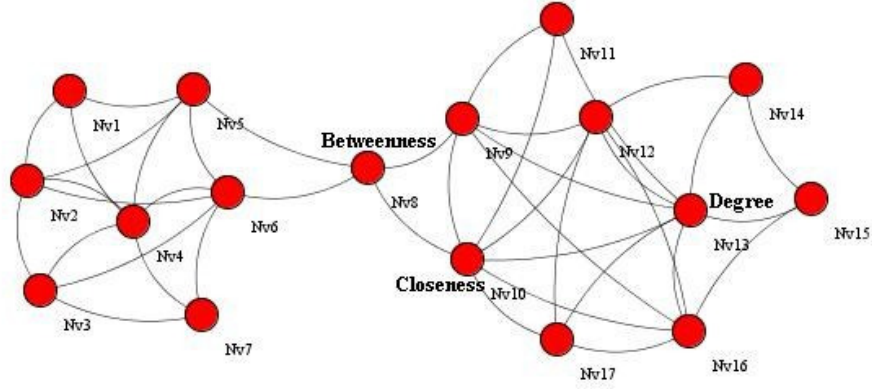


Figure 1: Example of which nodes have highest value of some centrality measures in a social network.

### 2.1.5.1 Degree Centrality

The first and simplest centrality measure is degree centrality. Degree centrality refers to the number of ties a node has to other nodes. Degree is a measure in some sense of the popularity of an actor. Actors that have more ties may have multiple alternative ways and resources to reach goals – and thus be relatively advantaged. If the network is directed, then we have to distinguish between indegree and outdegree. Indegree is the count of the number of ties directed to the node, and outdegree is the number of ties that the node directs to others. The outdegree centrality is defined as

$$C_{D_o}(i) = \sum_{j=1}^n a_{ij} \quad (2.1)$$

Where  $a_{ij}$  is 1 in the binary adjacency matrix  $A$  if an edge from node  $i$  to  $j$  exists, otherwise it is 0. Similarly, the indegree centrality is defined as

$$C_{D_i}(i) = \sum_{j=1}^n a_{ji} \quad (2.2)$$

Where  $i$  describes the node  $i$  and  $a_{ij}$  is 1 if an edge from node  $j$  to  $i$  exists, otherwise it is 0. The degree measure for a node focuses only on its directly connected neighbour nodes. To get the global position of a node, this algorithm is not applicable. Additionally, it assumes the more ties a node has, the more important it is. However, it does not consider the relative importance of its connected neighbour nodes. Therefore, other centrality measures must be used to find the global position of a node within the network.

### 2.1.5.2 Closeness Centrality

Recall earlier I have introduced the concepts of walk, path and geodesic distance in graph theory. A geodesic path is the shortest path, in terms of number of edges traversed, between a specified pair for nodes. It is often the optimal or most efficient connection between two nodes. The shortest-path distance between two vertices is defined as the length of a geodesic path that connects them.

Closeness is a measure of the degree to which an individual is near all other individuals in a network. Closeness centrality measures how close a node is to all other nodes in the graph. The closeness centrality of a node in a graph is the inverse of the average shortest-path distance from the node to any other node in the graph. It is defined as the mean geodesic distance between a node and all other nodes reachable from it. Moreover, closeness can be regarded as a measure of how long it will take information to spread from a given node to other reachable vertices in the network. The closer a node is to all other nodes, the easier information may reach it, the higher its centrality. Formally the closeness centrality is defined as

$$C_c(i) = \frac{(n-1)}{\sum_{j=1}^n d(i, j)} \quad (2.3)$$

Where  $d(i, j)$  denotes the distance between node  $i$  and  $j$ , which is the shortest path between node  $i$  and  $j$ . The idea behind this measure is that a node's closeness is measured by how quickly it can interact with all other nodes in the network. The important actors considered as central are often productive in communication with all other actors. One problem with this definition of closeness centrality is that it cannot be used for disconnected graphs in which some of the  $d(i, j)$  will be  $\infty$  and the equation (2.3) will be undefined. Obviously, this algorithm is much more time consuming than the degree centrality algorithm since it needs to calculate the geodesic distance between pair of nodes.

### 2.1.5.3 Betweenness Centrality

Another important class of centrality measures is the class of betweenness measures. The idea is that a node is central if it lies between other nodes on their geodesics

path, implying that, in order to have a large betweenness value, the node must be between many of nodes via their geodesics path. In fact, interactions between two non-adjacent nodes might depend on the other nodes in the set of nodes, especially those nodes that lie on the path between the two. Those nodes can therefore control the interaction between the two non-adjacent nodes. In many social contexts an actor with high betweenness will exert substantial influence by virtue not of being in the middle of the network but of lying “between” other actors in this way.

The simplest and most widely used betweenness measure is that of Freeman (1979). Formally, the betweenness of node  $i$  is defined to be the fraction of shortest paths between pairs of nodes in a network that pass through  $i$ , thus the betweenness centrality is defined as

$$C_B(i) = \sum_{s < t} \frac{g_{st}(i)}{g_{st}} \quad (2.4)$$

Where  $g_{st}(i)$  is the number of shortest paths linking the two nodes  $s$  and  $t$  that path through  $i$ , and  $g_{st}$  is the total number of shortest paths from  $s$  to  $t$ . This may be normalised by dividing through the number of pairs of nodes not including  $i$ , which is  $(n-1)(n-2)$  for directed graphs and  $(n-1)(n-2)/2$  for undirected graphs.

Meanwhile, betweenness centrality is an important indicator of control over information exchange or resource flows within a network. A node that is situated on the geodesics between many pairs of nodes is very important to the flow of information within the network. In a network in which flow is entirely or at least mostly along geodesic paths, the betweenness of a node measures how much flow will pass through that particular node.

The betweenness centrality measurement offers the most general and powerful way of determining the importance of actors. However, it assumes that actor  $i$  only chooses the shortest path to transmit information to actor  $j$ , and the probability of choosing multiple shortest paths is evenly distributed. It simply omits the fact that an actor which is on geodesic and has large degree may be more expansive than any other actors who are also on geodesic but have a relatively small degrees. The actor with a large degree should have a higher probability to be chosen on the geodesic.

Additionally, it omits all paths between  $i$  and  $j$  which have longer length than the shortest paths. This assumption is not necessarily valid. For example, in real world internet connection, it is possible for one computer to choose a path other than the shortest path to send messages to another computer.

The computation of betweenness centrality is computationally expensive. Brandes(2001) proposed an algorithm for betweenness that exploits the sparseness of typical networks to reduce running time from  $O(n^3)$  to  $O(n^2 + nk)$  and memory consumption from  $O(n^2)$  to  $O(n + k)$ . Moreover, other shortest-path based measures, like closeness, can be computed simultaneously within the same bounds.

#### 2.1.5.4 Eigenvector Centrality

The eigenvector centrality is another measure of the importance of a node in a network. This measure is a little more complicated than those considered previously and it is usually defined as the limits of some iterative process. The core idea behind the eigenvector centrality is that an important node is connected to important neighbours. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people. Being nominated as powerful by someone seen by others as powerful should contribute more to one's perceived power.

Typically the eigenvector centrality scores of a network's nodes were determined from the entries of the principle eigenvector of the network's adjacency matrix. Formally, the centrality of node  $i$  is a function of the centrality of the nodes connected to  $i$ . Let  $A$  be the binary adjacency matrix of the network and be the principal eigenvector corresponding to the maximum eigenvalue  $\theta$ . The eigenvector centrality for a node  $i$  can be defined as a single element of the eigenvector calculated as:

$$C_E(i) = x_i = \frac{1}{\theta} \sum_{j=1}^n a_{ji} x_j \quad (2.5)$$

Here, each actor's status is merely proportional to the weighted sum of the actors to whom he is connected. The eigenvector centrality defined in this way accords each

vertex a centrality that depends on both on the number and the quality of its connections: having a large number of connections still counts for something, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts (Newman 2007). This centrality measure turns out to be a revealing measure in many situations. For example, a variant of eigenvector centrality is employed by the well-known web search engine Google to rank Web pages, and works well in that context.

## **2.1.6 Authority Ranking Measures**

Most of the previously described centrality measures except eigenvector centrality disregard the type of node. There are very influential vertices to which a connection is more valuable than to others. With regard to social networks, a connection to a node with high centrality might be more valuable than to a node with only one neighbour. Web search algorithms such as HITS (Kleinberg 1999) and PageRank (Page, et al. 1998) leverage this information by using the link structure of a network of the webpages to assign weights to each page in the network. The weights can then be used to rank the pages as authoritative sources. These algorithms share a common underpinning; they find a dominant eigenvector of a non-eigenvector matrix that describes the link structure of the given network and use the entries of this eigenvector as the page weights (Farahat, et al. 2006). HITS emphasizes mutual reinforcement between authority and hub webpages, while PageRank emphasizes hyperlink weight normalization and web surfing based on random walk models (Ding, et al. 2002). Below I will discuss each algorithm.

### **2.1.6.1 HITS Algorithm**

Closely related to eigenvector centrality is a web search algorithm called HITS (Hyperlink-Induced Topic Search) introduced by Kleinberg (1999). The premise of the algorithm is that a web page serves two purposes: to provide information on a topic and to provide links to other others giving information on a topic. This gives rise to two ways of categorizing web pages: authoritative pages and a set of hub pages. Authoritative pages are those web pages pointed by a large number of hubs and hubs are those web pages that point to a large number of authoritative pages. The HITS algorithm is an iterative algorithm developed to quantify each page's value as



an authority and as a hub. For each page  $i$  a nonnegative authority weight  $C_A(i)$  and a nonnegative hub weight  $C_H(i)$  is associated, and the mutually reinforcing relationship between authorities and hubs is represented as

$$C_A(i) = \sum_{j=1}^n a_{ji} C_H(j) \quad (2.6)$$

$$C_H(i) = \sum_{j=1}^n a_{ij} C_A(j) \quad (2.7)$$

Where  $a_{ji}$  is 1 if an edge from node  $j$  to  $i$  exists otherwise 0. An iterative algorithm is used to find the equilibrium values for the authority and hub weights of a web page or node in a network respectively. This is reached if the difference of the weights between two iterations is less than a threshold value. When the equilibrium is reached, the most central nodes are those with the highest authority weight.

### 2.1.6.2 PageRank

Another popular web search algorithm is the PageRank algorithm introduced by Page and Brin(1998). PageRank uses an idea similar to HITS that a “good” webpage should connect to or be pointed to by other “good” webpages. However, instead of mutual reinforcement, it adopts a web surfing model based on Markov process in determining the scores. Unlike HITS, it maintains only a single metric for each web page. The so called PageRank is transmitted from the source page to the link target, and the size of the contribution depends on the PageRank of the source page. So a link from a page that has large PageRank, such as the Yahoo home page, contributes more a link from a page with low PageRank. For example, if a web page has a link off the Yahoo home page, it may be just one link but it is very important one. This page should be ranked higher than many pages with more backlinks but from obscure places.

The PageRank of page or node  $i$  is the sum of contributions from its incoming links or edges. A constant damping factor  $d$  is the probability at each page that the “random surfer” will get bored and requests another random page. Additionally  $(1-d)$  is added to each node. This is done because if a node has an out-degree of zero then his PageRank would be zero. This zero-value would be passed down to the

original node. To avoid this, a constant value is added to the PageRank. The PageRank can be defined as:

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)) \quad (2.8)$$

Where page  $A$  has pages  $T_1 \dots T_n$  which point to it,  $C(A)$  is defined as the number of links going out page  $A$ , and the parameter  $d$  is a damping factor which can be set between 0 and 1.  $d$  is usually set to 0.85.

### 2.1.7 Clustering & Community Detection

Consider the social network of friendships or other acquaintances between individuals; it is a matter of common experience that such networks seem to contain dense pockets of people who “stick together”. In social science, such subgroups are typically called communities or clusters. Intuitively, a community is a cohesive group of nodes that are connected more densely to each other than to the nodes in other communities, within which node-node connections are dense, but between which connections are less dense (Porter, Onnela and Mucha 2009). Community structure is ubiquitous in social networks. A community in a social network might indicate a circle of friends, a community in the World Wide Web might indicate a group of pages on closely-related topics, and a community in a cellular or genetic network might be related to a functional module. Below in Figure 2 is an example of community structure in a social network.

Community structure is an important network property and can reveal many hidden features of the given network. The ability to detect community structure in a network could clearly have practical applications. Individuals belonging to the same community are probable to have properties in common, perhaps by interest or background. The communities in the blogspace often correspond to topics of interests. Monitoring the aggregate trends and opinions revealed by these communities provides valuable insight to a number of business applications, such as marketing intelligence and competitive intelligence. In biochemical or neural networks, communities may be functional groups, and separating the network into such groups could simplify functional analysis considerably. Hence, identifying the communities is a fundamental step not only for discovering what makes entities

come together, but also for understanding the overall structural and functional properties of large network (Du, et al. 2007).

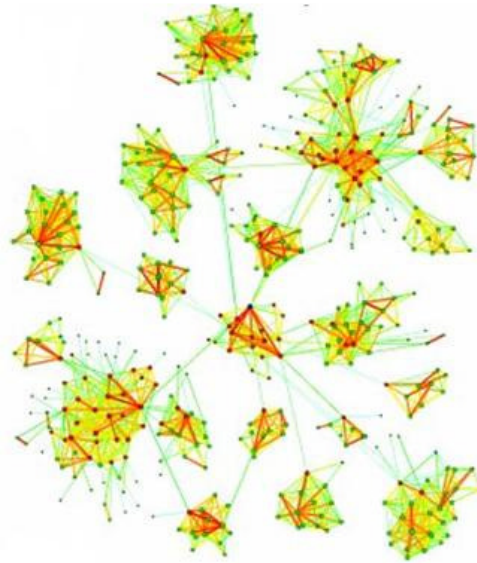


Figure 2: An example of community structure in a social network.

Community detection is closely related to the cluster analysis in data mining. In the classic clustering theory, entities are embedded in metric spaces and their similarity is derived from the distances between them. Thus, all pairwise similarities were known at the beginning. In social network analysis, the networks are typically sparse. Therefore, tailored concepts and algorithm are needed which take advantage of this special setting. Network data tends to be “discrete”, leading to algorithms using the graph property directly. There are numerous alternative methods for detecting communities in networks. These include hierarchical clustering (Navarro, Frenk and White 1997), partitioning graphs to maximize quality functions such as network modularity, k-clique percolation (Derényi, Palla and Vicsek 2005) and etc. The differences between many community-detection methods ultimately come down to the precise definition of more densely and the algorithmic heuristic followed to such sets (Porter, Onnela and Mucha 2009).

A popular quantitative definition called network modularity (Newman 2006), proposed by Girvan and Newman, is widely used as a quality metric for assessing the partitioning of a given network into communities. Good partitions, which have high values of the modularity, are those in which there are dense internal connections between the nodes within communities but only sparse connections between

different communities. The search for the largest modularity value is a NP-hard problem due to the fact that the space of possible partitions grows faster than any power of the system size. By definition, modularity considers the number edges which is smaller than expected:

$$Q = (\text{number of edges within communities}) - (\text{expected number of such edges}).$$

Most actual networks are made of highly overlapping cohesive subgroups of nodes simply because individuals often belong to numerous different kinds of relationships simultaneously (Palla, et al. 2005). For example, each of us may participate in many social cycles according to our hobbies, educational background, working environment and family relationships. As a result, when the network is large and the overlapping is significant, most of the existing algorithms in general will have high computation cost due to their heuristic optimization strategies (Du, et al. 2007).

### **2.1.7.1 Density, k-core & Cliques**

Before formally describing different community detection techniques, I start with some of the traditional techniques of detecting cohesive subgroups in social science. Intuitively, cohesion means that a social network contains many ties. More ties between people yield a tighter structure, which is, presumably, more cohesive. In network analysis, the density of a network captures this idea. It is the percentage of all possible lines that are present in a network. Maximum density is found in a complete simple network, that is, a simple network in which all pairs of vertices are linked by an edge or by two arcs, one in each direction.

K-core in graph theory was introduced by Seidman(1983) in 1983 and by Bollobas(1984)in 1984. Cores are clusters of nodes that are tightly connected because each node has a particular minimum degree within the cluster. These clusters are called k-cores and k indicates the minimum degree of each node within the core; for instance, a 2-core contains all nodes that are connected by degree two or more to other nodes within the core. A k-core identifies relatively dense subnetworks, so they help to find cohesive subgroups.

Cliques are a stricter structure form of a cohesive subgroup. A clique is a set of nodes in which each node is directly connected to all other nodes. In other words, a clique is a subnetwork with maximum density. The size of a clique is the number of nodes in it. Below in Figure 3 are cliques of size 3, 4 and 5. Maximal complete subnetworks of size 1 and 2 exist, but they are not very interesting because they are single nodes and edges, respectively. Therefore, cliques must contain a minimum of three nodes. Unfortunately, it is very difficult to identify cliques in large networks: the computational method is very time-consuming and even medium-sized networks may contain an enormous number of cliques. In social network analysis, one is typically interested in either a maximum clique or an enumeration of all maximal cliques. Despite the rather simple structure of cliques, both problems are NP-hard to solve.

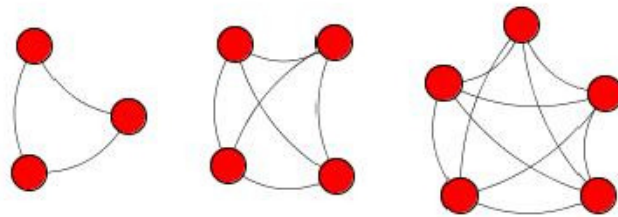


Figure 3: Cliques of Size 3, 4 and 5.

### 2.1.7.2 Edge Betweenness Clustering

A commonly used algorithm for communities finding is based on the graph-theoretic measure betweenness proposed by Girvan and Newman (2002). Rather than constructing communities by adding the strongest edges to an initially empty node set, they construct them by progressively removing edges from the original graph. As I have introduced in the previous section, for any node, node betweenness is defined as the number of shortest paths between pairs of nodes that run through it. It is a measure of the influence of a node over the flow of information between other nodes, especially in cases where information flow over a network primarily follows the shortest available path. Girvan and Newman have extended this definition to the case of edges, defining edge betweenness as the number of shortest paths that traverse it.

The main idea behind the algorithm is that if a network contains communities or groups that are only loosely connected by a few intergroup edges, then all shortest paths between different communities must go along one of these few edges. Thus, the edges connecting communities will have high edge betweenness. By iteratively removing these edges, groups can be separated from one another and the underlying community structure of the graph can be revealed. The edge betweenness distinguishes inter-community edges, which link many vertices in different communities and have high betweenness, from intra-community edges, whose betweenness is low. The algorithm proposed for identifying communities is simply stated as follows:

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweenness for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.

This algorithm has been employed by a number of authors in the study of such diverse systems as network of email messages, social network of animals and collaboration of jazz musician (Newman 2004). It also has been shown to have useful biological applications, in defining subsystems within food webs and to investigate biological function in protein-protein interaction networks (Dunn, Dudbridge and Sanderson 2005). However, as pointed out by Newman and Girvan, the principle disadvantage of their algorithm is the high computational demands it makes. In its simplest and fastest form, it runs in worst-case time  $O(m^2n)$  on a network with  $m$  edges and  $n$  vertices, or  $O(n^3)$  on a sparse network. With typical computer resources available at the time of writing, this limits the algorithm's use to networks of a few thousand vertices at most, and substantially less than this for interactive applications.

### **2.1.7.3 Voltage Cluster**

Another method for communities finding is using a physics approach by considering the social network as an electric circuit with each edge having the same resistance (Newman 2004, Wu and Huberman 2004). Specifically, social networks are modelled using concepts from electrical circuits, voltage, current, conductance and

etc. Source node has +1 volts, sink has 0 volts and current will flow from source to sink. The weight of the edges can be represented by conductance. When comes to communities structure, physically thinking, because nodes inside a community are densely connected, their voltages tend to be close. A big voltage gap happens about halfway between two communities, where the edges are sparse and the local resistance is large.

One advantage of the algorithm is that it allows for the discovery of a community surrounding a given node without having to extract all the communities out of a graph. One volt is applied to the given node, and zero volts to some other randomly choose node. The voltages at all nodes are then calculated using Kirchhoff's Laws, and the nodes are split into two groups by sorting all the voltages, picking the median voltage, and splitting the nodes on either side of this median into two communities. The important idea is that the voltages can be calculated approximately using iterative methods requiring only  $O(V + E)$  time but with the quality of approximation depending only on the number of iterations and not on the graph size.

This algorithm runs in linear time, but this particular method needs a priori knowledge of the number of expected communities. The reason behind the speed of this method lies in its focus on communities themselves and not on their hierarchical structures. In contrast, Girvan and Newman's betweenness method detects not only the communities but also the complete hierarchy tree using much longer times. One defect of the method is that the number of communities wish to divide the graph into should be specified, a piece of information which one does not often have beforehand (Wu and Huberman 2004).

## 2.2 Diffusion Processes in Social Networks

With the explosive growth of online social networks and great potential of viral marketing, studies of the information diffusion over social networks have attracted more and more attention. Traditionally, studies on diffusion process in social networks have span across multiple disciplines including sociology (Rogers 1995) , (Granovetter 1978), marketing (Mahajan, Muller and Bass 1990), and epidemiology (Anderson and May 1992). In social science and marketing literature there is a long history of research on the role of social network in new product diffusion. In the 1950s, Rogers' diffusion of innovation theory (Rogers 1995) has divided the innovation adoption process into different stage and classified the adopters of the innovation into categories. The two-step flow of communication model (Lazarsfeld, Berelson and Gaudet 1944, Katz and Lazarsfeld 1955) hypothesizes that ideas flow from mass media to opinion leaders and from them to a wider population. The influences stemming from the mass media first reach 'opinion leaders', who in turn, pass on what they read and hear to their friends for whom they are influential. In epidemiology, numbers of epidemic models have been proposed to describe the transmission of infectious disease among large populations. Recently, with the availability of large-scale network data, researchers have contributed a plethora of studies, approaches and theoretical contributions related to various aspects of the diffusion phenomenon.

My reasons to conduct a thorough review on this area are twofold. Firstly, despite the piles of studies related to diffusion processes in recent years, there are only rare research articles consolidating and reviewing the various options. There are many options and approaches. This makes it very difficult for researchers to understand the various approaches and identify their advantages and challenges. Second, this review can provide a good foundation and theoretical background on my work on modelling diffusion processes and empirically assess the phenomenon.

In the following sections, I aim to contribute an overview of the most prominent approaches related to the studies of the diffusion phenomenon. I present a framework and research overview for this area. I believe my framework can assist researchers



and practitioners to understand the challenges in studying information diffusion over social networks and identify suitable solutions to address those challenges.

### 2.2.1 Modelling Diffusion Processes

Kleinberg (2008) argues that while the outcomes of the diffusion processes are easily visible, their inner workings have remained elusive. According to Cha, Mislove and Gummadi(2009)neither the characteristics of the information diffusion nor the mechanisms by which information is exchanged are well understood. Models of diffusion process aim to provide a clear picture of how such dynamic process unfolds by defining the way information flows between individuals. From Bass model (Mahajan, Muller and Bass 1990)describing the new product diffusion to epidemic models modelling the spread of infectious diseases over large population, many attempts have been made to model such dynamic process from diverse areas.

I consider a collection of probabilistic and game-theoretic diffusion models. I categorize three major challenges faced when modelling the diffusion phenomenon: modelling the dynamic adoption process, the factors affect adopting decisions and the diffusion of competing technologies, as illustrated in Table 3. I then summarize common approaches to address these challenges: the adoption process can be described explicitly representing the step-by-step dynamics of adoption, as a coordination game or a markov process. Besides social influence, there are some other factors such as the intrinsic tendency to adopt might also affect people’s adopting decision. In real world scenario, it’s common that several competing technologies diffuse in the system at the same time. In the following sections, I address each challenges and approaches in detail.

Table 3: Diffusion modelling challenges and approaches.

<b>Challenge</b>	<b>Modelling Approaches</b>		
The dynamics of adoption process	Explicitly represent step-by-step dynamics of adoption	Game-theoretic approaches	Markov process
Modelling social influence	Threshold-rule	Infection rate	Individual heterogeneity

Modelling diffusion of competing technologies	Coordination game	Consider only one technology diffuse
---	-------------------	--------------------------------------

### 2.2.1.1 The Dynamics of Adoption Process

One of the most common approaches of modelling the adoption process is to explicitly represent the step-by-step dynamics of adoption (Kempe, Kleinberg and Tardos 2003). Typically, it assumes the dynamic process unfolds in discrete time unit with each individual following certain rule when making adopting decision. An individual who has adopted the behaviour is called being active and inactive otherwise. A set of individuals are chosen to be initial active set which corresponds to the early adopters of the behaviour. Starting with the initial active set the process then unfolds as follows: at each time step, individuals who were active at previous time step remain active; an individual will be activated according to certain rules such as a threshold number of his neighbours have been activated. The dynamic process continues until no more activation is possible. Below I use Figure 4 to illustrate this approach.

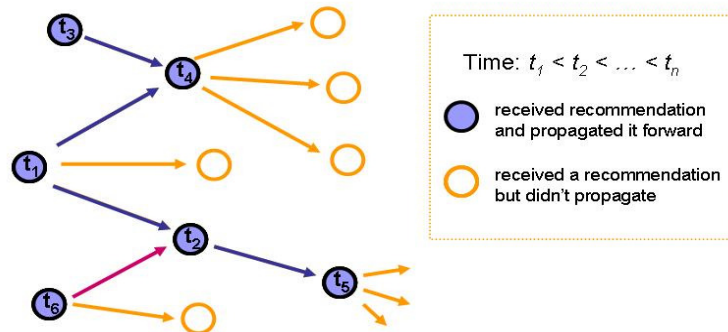


Figure 4: Explicitly represent the step-by-step dynamics of adoption process.

In the economics literature, diffusion models have been studied from a game-theoretic perspective (McPherson, Smith-Lovin and Cook 2001). This approach builds on work investigating how a new technology A might be spread through a social network of individuals who are currently users of technology of B. The diffusion process is modelled as the dynamic of a coordination game played on the social network, in which the adoption of a common strategy between players has a higher payoff. In particular, in every time step, each player in a social network has two available choices A and B. Each player receives a positive payoff for each of his

neighbours that has the same choice as he does, in addition to an intrinsic benefit that he derives from his choice (Gruhl, et al. 2004).

Beyond above approaches explicitly modelling the evolution of the actions over time, an alternative is to consider the diffusion process as a Markov process. Song et al. (2007) proposed a rate-based information flow model based on Continuous-Time Markov Chain. Each state in the Markov chain corresponds to one particular system configuration of all individuals. Each individual can be in one of two states (active or inactive), the weight is represented as the transition probability, and the delay is represented as the staying time in each state.

### **2.2.1.2 Modelling Social Influence**

An underlying premise many diffusion models build on is that people are influenced by their neighbours in the social network when making adopting decision. Social influence determines to a large extent what we adopt and when we adopt it. One simple way to capture this effect is to assign each individual in the network a threshold value (Granovetter 1978, Kempe, Kleinberg and Tardos 2003). An individual becomes active if a certain pre-specified number or fractions of his neighbours are active. The threshold value indicates the personal tendency of an individual to adopt the behaviour when his neighbours do. In addition to the number of adopted friends, how those friends are connected to one another could also have an impact on individual's propensity of adopting. It is argued that if two actors related to the same individual are also related to each other, they have greater power over that individual than if they were unrelated (Burt 2005). A primitive approach to incorporate this idea is to postulate the adoption likelihood of individual increases as a function of the density of relationships among their adopted neighbours (Katona, Zubcsek and Sarvary 2009).

Another way to encode neighbour influence is by using infection rate, inspired by the epidemic models. In the Independent Cascade Model (Kempe, Kleinberg and Tardos 2003, Goldenberg, Libai and Muller 2001) every time an individual contacts with an active neighbour, he has a constant chance of getting activated. Obviously, a constant infection rate seems not accurate enough. Kempe, Kleinberg and Tardos (2005) refined the Independent Cascade Model to interpret the idea that an

individual's receptiveness to influence depends on the past history of interactions with his neighbours. Contrary to a constant rate, an individual's probability of being activated is a function of the set of neighbours have already tried and failed to influence him. This modification is generally consistent with Leskovec, Adamic and Huberman's finding(Leskovec, Adamic and Huberman 2007)that the probability of infection decreases with repeated interaction when analyzing the patterns of influence in a recommendation network.

It is commonly accepted that social influence affects adopting decision. However, it is not the only factors that drive information diffusion. Van den Bulte and Stremersch(2004)argue that S-shaped diffusion curves can also result from heterogeneity in the intrinsic tendency to adopt. In the context of new product diffusion, individuals are different with respect to subject matter expertise, strength of opinion, personality traits, media exposure or perceived adopting costs. Efforts have been made to capture this effect. For instance, Katona, Zubcsek and Sarvary(2009)model the adoption decision of individuals as a binary choice affected by three factors: the local network structure formed by already adopted neighbours, the average characteristics of adopted neighbours, and the characteristics of the potential adopters. Hartline, Mirrokni and Sundararajan(2008)model a buyer's decision to buy an item is influenced by the set of other buyers that own the item and the price at which the item is offered.

Meanwhile, there is a debate in the literatures as to what role influential individuals actually play in social cascades. Duncan Watts have performed a number of interesting experiments which suggest that the role of influential is overstated (Watts and Dodds, 2007).

### **2.2.1.3 Modelling Diffusion of Competing Technologies**

Most models discussed above typically focus on the diffusion of one behaviour or technology. What is often ignored is that several different behaviours may coexist in a system at the same time and possibly compete with each other. Some may have a better chance to survive and spread than others. In fact, this scenario frequently arises in the real world. According to (Carnes, et al. 2007)in consumer market, producers of consumer technologies often must introduce a new product into a

market where a competitor will offer a comparable product, which makes them vie for sales with competing word-of-mouth cascades. Thus it is important to examine the diffusion of competing technologies in social networks.

As mentioned earlier, such phenomena can be modelled as a coordination game played on the edges of the social network with multiple equilibria. An influential paper by Morris (2000) provided a set of elegant game-theoretic characterizations for when these qualitatively different types of equilibria arise in terms of the underlying network topology and the quality of technology A relative to technology B. In recent work, Immorlica et al. (2007) incorporates compatibility between technologies and discuss how this effects the diffusion. Their results show that in some cases, for one technology to survive the introduction of another, the cost of adopting both technologies must be balanced with a narrow, intermediate range. Another work by Carnes et al. (2007) assumes that consumers will use only one of the two products. They propose two models for the spread of influence of competing technologies and consider the influence maximization problem from the follower's perspective.

### **2.2.2 Empirical Diffusion Studies**

Several recent studies have empirically measured the dynamics of information diffusion through large-scale social networks such as blogspace (Gruhl, et al. 2004), person-to-person recommendation network (Leskovec, Adamic and Huberman 2007) and mobile phone network (Onnela, et al. 2007). These studies have offered some deep insights and fundamentally advanced our knowledge of how information diffuses in social networks. On the other hand, contagion and cascading behaviour have been employed in proposals for social computing applications such as word-of-mouth recommendation systems. In this section, I provide an overview of recent research examining the diffusion phenomenon empirically.

Most of the recent empirical studies on information diffusion focus on addressing the following fundamental questions:

- What are the characteristics of social influence?
- What are the patterns of information cascade?
- How does network structure affect diffusion process?

### 2.2.2.1 Characteristics of Social Influence

If we view the diffusion process as a cascade of social influence, a natural starting point is to understand the local mechanism of the influence. Social influence can be described as the actions of an individual can induce his or her friends to behave in a similar way. A number of recent diffusion studies (Leskovec, Adamic and Huberman 2007, Backstrom, et al. 2006, Christakis and Fowler 2007, Anagnostopoulos, Kumar and Mahdian 2008, Crandall, et al. 2008) have measured to what extent social influence affects adopting decision empirically.

Backstrom et al. (2006) investigated the membership problem in online communities and measured how propensity of individuals to join a community depends on friends already within the community. Specifically, they measured the joining probability as a function of the number of friends already in the community. As shown in Figure 5, the adoption curve exhibits a diminishing returns pattern in which it continues increasing, but more and more slowly, even for large numbers of friends. In addition, they examined the dependence of joining tendency on more subtle features such as how these friends are connected and found that joining probability increases as the density of linkage increases among the individual's friends in the community.

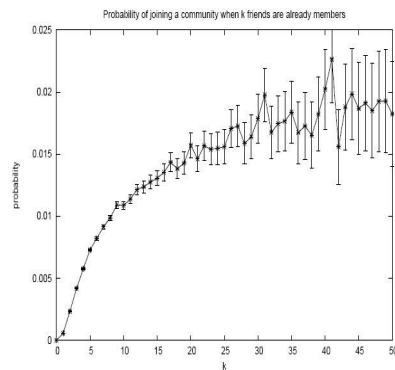


Figure 5: Probability of joining a community when k friends are already members.

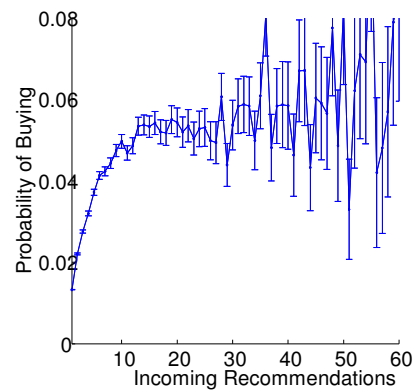


Figure 6: Probability of buying DVD with incoming recommendations

This kind of diminishing returns pattern has also been observed in many other studies. For example, Christakis and Fowler (2007) studied the spread of obesity in large social network over 32 years and found out that people were most likely to become obese when friends became obese. Leskovec et al. (2007) also found that the probability of purchasing a product increases with the number of recommendations

received, but quickly saturates to a constant and relatively low probability, as shown in Figure 6. However, Kleinberg (2007) argued that the dependence of adopting probability on the number of friends adopted expressed in this way reflects an aggregate property of the full population, and does not imply anything about an particular individual's response to their friends' behaviour. One is observing social activity in aggregate and doesn't necessarily know what any one particular individual or social connection signifies (Kleinberg 2008).

Anagnostopoulos et al. (2008) argues that while these studies have established the existence of correlation between user actions and social affiliations, they do not address the source of the correlation. There are factors such as homophily (McPherson, Smith-Lovin and Cook 2001) or unobserved confounding variables that can induce statistical correlation between the actions of friends in a social network. Homophily is the tendency for people to choose relationships with people who are similar to them, and hence perform similar actions. Distinguishing social influence or interpersonal induction from homophily requires dynamic, longitudinal network information such as the emergence of ties between individuals (Christakis and Fowler 2007). Recently, Anagnostopoulos et al. (2008) proposed a statistical test to distinguish social influence from correlation using time series data of user actions. Crandall et al. (2008) developed techniques for identifying and modelling the interactions between social influence and social selection process and found an elaborate interplay between the two factors.

### **2.2.2.2 Patterns of Information Cascade**

Although above studies have shed light on the mechanisms of social influence, the overall patterns by which the influence spreads through social networks have been a mystery. Several recent studies have been conducted to illustrate the existence of cascade and observe patterns of cascading behaviour. In particular, Leskovec et al. (2007) consider information cascades in a recommendation network. According to their observation, the distribution of cascade sizes can be approximated by a heavy-tailed distribution. Generally cascades are shallow but occasional large bursts also occur. Another recent study on cascading behaviour in large blog network (Leskovec, et al. 2007) found that blog posts do not have a bursty behaviour; they only have a weekly periodicity.

Liben-Nowell and Kleinberg (2008) traced the information cascade process on a global scale by using methods to reconstruct the propagation of massively circulated Internet chain letters. Contrary to predictions that large-scale information spreads widely and reaches many people in very few steps, their results show that the progress of these chain letters proceeds in a narrow but very deep tree-like pattern, continuing for several hundred steps.

### **2.2.2.3 Role of Network Structure**

Researchers have long emphasized the important role played by the network structure in determining properties of information diffusion. However, the way such dynamic process is affected by network structure is still poorly understood. How widely does information spread? Does it spread only in local region? Does it spread quickly on a dense network? Several studies on real-world network data have been conducted to address questions like these.

Typically strength of the influence one individual exerts on another is represented by the strength or weight of the tie connecting them. Tie strength varies from strong to weak depending on the number and types of resources they exchange, the frequency of exchanges, and the intimacy of the exchanges between them. In the case of a telecom call network, tie strength can be defined by the aggregated duration or frequency of calls exchanged between two individuals over observing period. Strong ties bear greater influence on the receiver's behaviour than weaker ties due to the frequency and perceived importance of social contact among strong-tie individuals.

Granovetter's weak ties hypothesis (Granovetter 1973) states that weak ties typically act as connectors between different communities or circles of friendship. Using mobile call records, Onnela et al. (2007) have observed a coupling between interaction strengths and the network's local structure, confirming the weak tie hypothesis. Specifically, they found that weak ties appear to be crucial for maintaining the network's structural integrity, but strong ties play an important role in maintaining local communities. In addition, they investigated how the dynamics of different tie strengths influence the spread of information in the network. They show that the coupling between tie strength and network structure significantly slows the diffusion process, resulting in dynamic trapping of information in communities



and find that both weak and strong ties have a relatively insignificant role as conduits for information.

### **2.2.3 Diffusion Application**

Understanding how information diffuses on a large scale may have applications in a variety of areas. One example is viral marketing. Consider the following scenario: A company tries to use word-of-mouth effects to market a product with limited budget. How should they choose the set of customers to target? Influence maximization (Kempe, Kleinberg and Tardos, 2003) is the problem of identifying influential set of nodes in a network that can trigger the largest cascade. This problem was first introduced by Domingos and Richardson and has been studied extensively from a theoretical perspective. Kempe et al. (2003) has proven this problem is NP-hard and provided a simple approximate solution with provable performance guarantees. Hartline et al. (2008) focus on the algorithmic question of finding revenue maximizing marketing strategies. They identify a family of influence-and-exploit strategies that are easy to implement and approximately optimal.

### **2.3 Social Network Analysis Tools**

Along with the increased relevance of social network analysis and the growing size of considered networks, adequate software for social network analysis is becoming more and more important. Social network analysis tools are used to identify, represent, visualize or simulate nodes (e.g. agents, organizations, or knowledge) and edges (relationships) from various types of input data (relational and non-relational), including mathematical models of social networks. Network analysis tools allow users to investigate representations of networks of different size - from small (e.g. families, project teams) to very large (e.g. the Internet, disease transmission). These tools often feature an impressive number of analysis techniques that users can perform on networks. More specifically, they can help users to answer questions like: what patterns are created by the aggregate of interactions in a network? How are the actors connected to one another? What social roles exist and who plays critical roles like connector, broker and etc.? How do network structure correlate with the contributions people make within the social network?

Visual representations of social networks are important to understand network data and convey the result of the analysis. Visualization is often used as an additional or standalone data analysis method. For example, algorithmically generated layouts have useful spatial properties: a force-directed layout can be quite effective for spatially grouping connected communities, while a radial layout intuitively portrays network distances from a central actor. Color, size, and shape have been used to encode both topological and non-topological properties such as centrality, categorization, and gender. With respect to visualization, network analysis tools are used to change the layout, colors, size and other properties of the network representation.

Over the years there are dozens of software tools designed to help analysts understand social networks. For example, Pajek is a network analysis and visualize program, specifically designed to handle large data sets. GUESS is a novel graph exploration system that combines an interpreted language with a graphical front end (Adar 2006). JUNG is a Java toolkit that provides users with a framework to build their own social network analysis tools (O'Madadhain, et al. 2005). A systematic overview and comparison of a selection of software packages for social network analysis was provided by Huisman and Van Duijn (2005). A large list of software packages and libraries can be found under Computer Programs for Social Network Analysis, maintained by the International Network for Social Network Analysis (INSNA). According to an overview of common social network analysis software platforms, SNA tools can be divided into the following broad categories as shown in Table 4:

Table 4: Summary of social network analysis tool types.

<b>Type</b>	<b>Description</b>
Advanced/Academic social network analysis tools	<p>Often used in academic settings and intended for the most sophisticated types of social network analysis</p> <p>Often built for performance as opposed to usability</p> <p>User guides and help files are not comprehensive or are written for more sophisticated audiences</p> <p>Example: Pajek, UCINET, GUESS.</p>

<p>Accessible but advanced social network analysis tools</p>	<p>Used in more general settings, including corporate environments</p> <p>Built with the user in mind and tend to be more intuitive and easier to use than tools for primarily academic applications</p> <p>Software help files are more comprehensive and user guides are written for a general user audience</p> <p>Example: NetMiner, Visone</p>
<p>Simple, easy to use social network analysis tools</p>	<p>Can be used by users less familiar with social network analysis</p> <p>Tools are built without complex functionality and are very easy to navigate and use</p> <p>Help files are simple and clear</p> <p>Example: Smart Network Analyzer</p>
<p>Online Tools that enable visualization of pre-existing user generated data</p>	<p>Used to analyze existing data made available by users</p> <p>Often simple to use with intuitive functionality</p> <p>Example: Xigi</p>

## 2.4 Network Visualization

Network visualization is the process of creating a useful visual representation of a social network. Visual representations of social networks are important to understand network data and convey the result of the analysis. Good visualization may reveal the hidden structure of the networks and amplifies human understanding, thus leading to new insights and new findings. Social network research has made extensive use of visualization since Moreno first introduced the sociogram (Freeman 2000, Brandes, Raab and Wagner 2001). Actors are usually represented as points, and relations among actors are represented by lines, with relational direction indicated by arrows. Network visualization deals with all aspects of representing relational structures and spans a diverse field ranging from matrix-like representations to figurative drawings of nodes connected by lines. Successful

visualization enables users to gain meaningful high-level information from an overview as well as to ascertain the details of each node and link.

### **2.4.1 Graph Layout Algorithm**

Nodes in a graph can be drawn anywhere in a simple visualization, including on top of one another. The basic graph drawing problem can be put simply: given a set of nodes with a set of edges, calculate the position of the nodes and the curve to be drawn for each edge (Herman and Marshall 2000). Graph layout try to position the nodes of that graph using a layout algorithm that attempts to fulfil certain aesthetic requirements; then add and remove control points of edges in the graph using a layout algorithm that attempts to fulfil certain aesthetic requirements. Exact what these aesthetic criteria are depend upon individual application or layouts requirements. Generally, these might involve spreading out vertices evenly without them overlapping each other, avoiding edges overlapping vertices and crossing other edges, clustering connected vertex neighbours and ordering vertices to reflect overall graph direction.

According to Shneiderman and Airs (2006), the literature on graph layout has been dominated by force-directed strategies because they produce elegant spreading of nodes and reasonable visibility of links. A second common layout strategy, which generates familiar and comprehensible layouts, visualizes nodes on a geographical map. Another common strategy uses a circular layout for nodes that produces an elegant presentation with crisscrossing lines through the centre of the circle. In the development of my tool I have selectively implemented some of the graph layout algorithms that are most common in practice. In the following section, I first examine the criteria of good layout then I introduce each of the graph layout algorithms in detail.

#### **2.4.1.1 Criteria of Good Layout**

Network visualization can be an important tool over and beyond the mere illustration of data, and that not every network visualization is equally effective in doing so (Brandes, Kenis and Raab 2006). According to Tufte's principles of graphical excellence (Tufte and Howard 1983) that good graphics is the well-designed

presentation of interesting data – a matter of substance, of statistics, and of design. Regarding to graph visualization, Brandes et al. (1999) argued that in order to produce effective visualizations one has to clearly identify the relevant information, that is, filter, transform, and process the collection of actors, links, and attributes to identify the interesting substance, define and appropriate mapping to a graphical representation, and generate the image accordingly without introducing artifacts. More specifically, they have defined three aspects which should be carefully considered when creating visualizations:

- The substance the viewer is interested in.
- The design which maps the data to graphical features.
- The algorithm employed to realize the design.

Graph drawing aesthetics are used to pinpoint the criteria that make a graph easier to perceive. In fact, even with small graphs, a bad layout can make a graph difficult to comprehend. While it is difficult to interpret the graph with a random layout in Figure 7, it becomes considerably easier to perceive the structure when its layout is improved as in Figure 8.

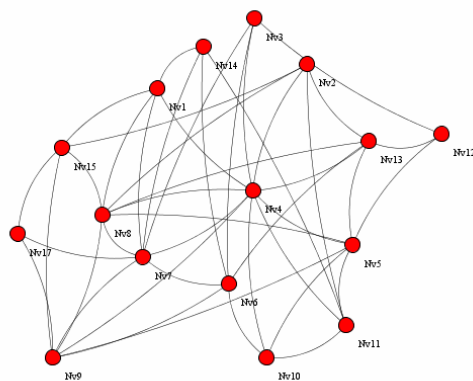


Figure 7: A graph with random layout.

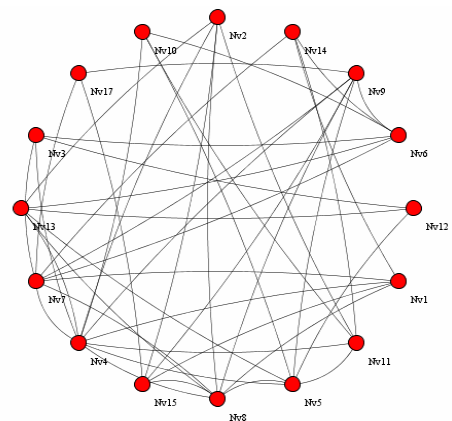


Figure 8: A graph with circular layout.

Graph layout algorithms typically take into account one or more aesthetic criteria, with the assumption that by doing so, the readability of the drawing is increased. The success of these algorithms is typically measured by their computational efficiency and the extent to which they conform to aesthetic criteria (Purchase, Carrington and Allder 2002). Some of the frequently used aesthetic criteria include, for example, minimising the number of edge crossings, maximising the depiction of symmetry,

and maximising the minimum angle between adjacent edges leaving a node (Purchase 2002).

#### **2.4.1.2 Circular Layout**

Circular layouts are among the most prominent and oldest conventions used to draw graphs. In such layouts, all nodes are drawn on the perimeter of a circle, while the edges connecting these nodes are line segments passing within the circle (Gansner and Koren 2006). Circular layouts are a very popular representation of social networks because the uniform placement induces no implicit hierarchy between the nodes, edges and nodes are clearly separated and do not cover each other, and last but not least they are easy to understand. It helps point out on the first sight which nodes are connected to many other nodes (graph density), and helps you find out in an intuitive way the critical nodes of your graph. In particular, a circular layout is appropriate for applications that emphasize the clustering decomposition of a graph, where each cluster is drawn on a separate circle (Gansner and Koren 2006). Note that a circular layout is also the base for the radial and group visualizations and, with extensions, for the multi-circular visualization.

In general, circular layout is a very simple layout algorithm which gives a good overview of the number of nodes and edges in a network. It can provide a compact presentation, focusing on individual nodes and edges. On the other hand, this strong regularity can obscure other information. For example, these drawings can be very dense, and following paths on them can be difficult. Circular layouts are a rather restrictive layout scheme, offering a simple and highly regularized picture of the graph where nodes cannot be occluded. The limiting nature of circular layouts makes it very important to capitalize on all available degrees of freedom (Gansner and Koren 2006). It is not useful for visualizing large networks and it does not give us any idea about the structure of the network. Typically it is used to visualize small and medium networks.

The algorithm is trivial and takes as input the number of nodes and the graph center position. It computes the adequate radius of the circle centred at the old graph center, nodes will be laid at its circumference. Once the radius is calculated, the algorithm

places the nodes according to an assigned angle ( $360/\text{number of nodes}$ ) and the circle radius.

### **2.4.1.3 Geographic Layout**

As I stated previously, another common layout strategy, which generates familiar and comprehensible layouts, visualizes nodes on a geographical map. In some network, the spatial information about the nodes may be available. Hence, the network may have a natural spatial layout as does a geographical trade-flow network, or may be abstract as in a personal communications network (Becker, Eick and Wilks 1995). Geographic maps are the most intuitive way to describe and explain the spatial organization of a phenomenon that involves geographically referenced data, that is to say data which has locational references within its structure.

### **2.4.1.4 Force-directed Algorithms**

Among various graph drawing techniques reported in the literature, the force-directed approaches have received much attention and have become very popular for drawing general undirected graphs. Force-directed approaches use a physical analogy to model the graph drawing problem, namely a system of forces acting on the nodes (Tunkelang 1999). It views the graph as a virtual physical system, where the nodes of the graph are bodies of the system. These bodies have forces acting on or between them. Often the forces are physics-based, and therefore have a natural analogy, such as magnetic repulsion or gravitational attraction (Quigley 2003). The most straightforward force-directed algorithm uses attractive forces between adjacent nodes and repulsive forces between all other pair of nodes (Tamassia 2006).

Regardless of the exact nature of the forces in the virtual physical system, force directed algorithms aim to compute a locally minimum energy layout of the nodes (Quigley 2003). This is usually achieved by computing the forces on each node and iterating the system in discrete time steps. By using certain law of physics, forces around each node is computed and then equilibrium state is sought. New values for the node position are computed and the whole net of nodes is rearranged. An energy model is associated with the graph layouts. Low energy states correspond to nice

layouts. Hence, the drawing algorithm is an iterative optimization process – energy minimization. A force directed graph drawing system consists of:

- Model: a force system calculating force based on the vertices and edges.
- Algorithm: a method for finding the equilibrium state of the force system, i.e. where the total force on each vertex is 0.

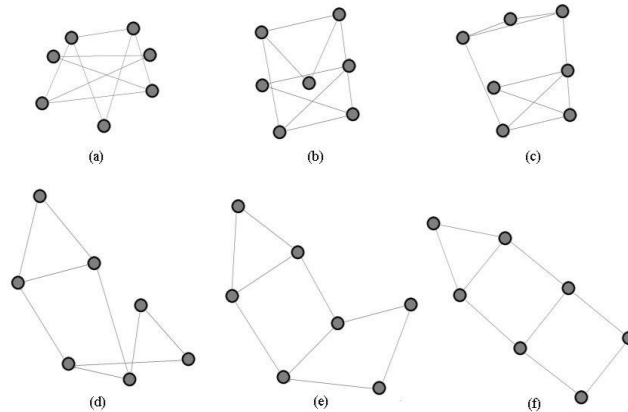


Figure 9: A graph drawing through a number of iterations of a force directed algorithm.

The trick is to find a force system that is both quick to calculate and forms a good graph layout. Here in Figure 9 is an example illustrates the step by step iteration of force directed algorithm, where the edges can be modelled as gravitational attraction and all nodes have electrical repulsion between them (Quigley 2003).

Force-directed algorithms are often used in graph drawing due to their flexibility, ease of implementation, and the aesthetically pleasant drawings they produce (Quigley 2003). Graphs drawn with these algorithms tend to be aesthetically pleasing, exhibit symmetries and tend to produce crossing-free layouts for planar graphs (Tamassia 2006). Additionally, they are very intuitive because of their relation to the real physical world. Furthermore, it is comparably easy to incorporate custom constraints into these models, e.g., to consider not only “points and lines” but also the area of the elements or varying preferred distances between different pairs of vertices.

However, classical force directed algorithms are unable to handle larger graphs due the inherent  $N^2$  cost at each time step, where  $N$  is the number of bodies in the system (Quigley 2003). This is a common problem and has prohibited the practical use of force directed algorithms for even moderately sized graphs of a few hundred



nodes. The utility of the basic force-directed approach is limited to small graphs and results are poor for graphs with more than a few hundred nodes. There are multiple reasons which traditional force-directed algorithms do not perform well for large graphs. One of the obstacles to the scalability of these approaches is the fact that the physical model typically has many local minima (Tamassia 2006). Even with the help of sophisticated mechanisms for avoiding local minima the basic force-directed algorithms are not able to consistently produce good layouts for large graphs. Besides, typically force directed algorithms produce a different final layout each time the algorithm is invoked, which is disorienting. Furthermore, when they are applied to graphs with labelled nodes, the resulting layouts suffer from severe node occlusions.

Generally speaking, force-directed layouts are particularly suited for sparse graphs with few shortcuts only, i.e., only some edges connect very different regions of the graph. Resulting layouts often expose the inherent symmetric and clustered structure of a graph, and feature a well-balance distribution of vertices and few edge crossings. Contrarily, in dense graphs or graphs of low diameter, vertices tend to cluster in the central area of the drawing.

#### **2.4.1.5 Fruchterman-Reingold Layout**

The earliest heuristics of force-directed placement were based on the spring embedder model (Eades 1984). Nodes are considered as mutually repulsive charges and edges as springs that attract connected nodes. The spring embedding algorithm assigns force between each pair of nodes. When two nodes are too close together, a repelling force comes into effect. When two nodes are too far apart, they are subject to an attractive force. This scenario can be illustrated by linking the vertices with springs—hence the name "spring embedding."

The spring system starts with a random initial state, and the vertices move accordingly under the spring forces. It iteratively improves an initial random layout by moving all vertices simultaneously in proportion to the net force acting on them. By iteratively computing the forces on all vertices and updating positions accordingly, the system approaches a stable state, in which no local improvement is possible. An optimal layout is achieved as the energy of the system is reduced to

minimal. One way to minimize the energy function is by iteratively moving each of the vertices along the direction of the spring force until an approximate equilibrium is reached. Multilevel techniques are used to overcome local minima. The equilibrium configuration, where the sum of repulsive forces due to rings and attractive forces due to springs is zero, normally results in a good drawing. Below in Figure 10 is an example of the spring embedder layout.

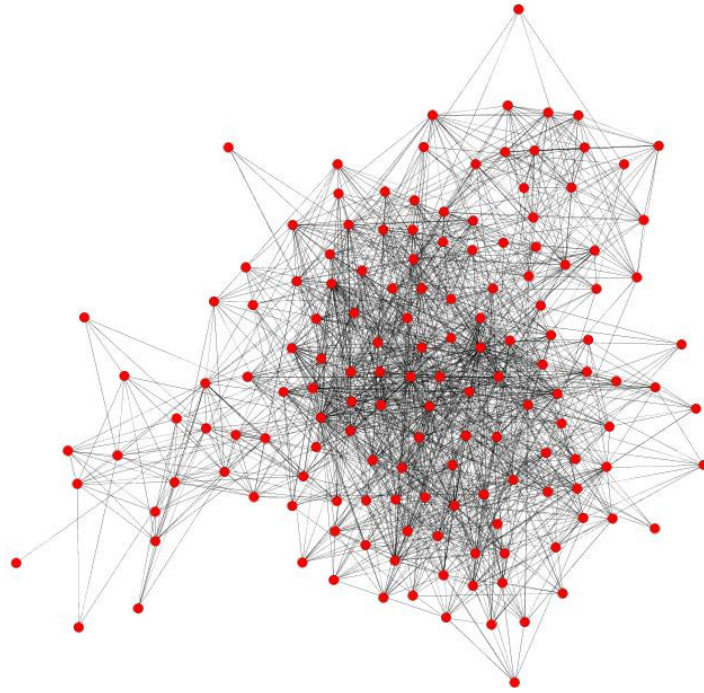


Figure 10: An example of the spring embedder layout.

Despite its simplicity, the spring embedder produces satisfactory output in many cases. To even out some shortcomings of the method, several refinements have been developed. These refinements mainly aim at faster computation, but sometimes also at improved quality of the layout. Fruchterman and Reingold(1991) subsequently presented an effective modification of the model. A number of heuristics is used by Fruchterman and Reingold to speed up many aspects of layout computation. Firstly, the forces are modified to allow faster evaluation. A second heuristic to speed up computation does not change the objective function, but the precision of evaluation. Their algorithm ignores the repulsive forces between vertices far away from each other and introduces a cooling parameter which increasingly limits vertex movement in later iterations. Additionally, the maximum displacement of each node in an iteration is limited by a constant that is slightly decreased with each iteration. These modifications introduce a trade-off between quality and computation time: the

algorithm converges faster to a stable state but this state is typically not a local optimum.

The Fruchterman-Reingold Algorithm (Fruchterman and Reingold 1991) is useful for visualizing very large undirected networks. It guarantees that topologically near nodes are placed in the same vicinity, and far nodes are placed far from each other. Using a global temperature creates an overall satisfying layout; however there will be deficiencies in some local areas of the graph. This can be improved by using the "adaptive temperature scheme" based on local temperatures.

#### **2.4.1.6 Kamada-Kawai Layout**

Another very popular algorithm fall into the category of force-directed is Kamada-Kawai (KK) layout (Kamada and Kawai 1989). It is also based on the idea of balanced spring system and energy minimization. However, different to FR-layout, KK-layout attempts to utilize the derivatives of the force equations to achieve faster convergence. Unlike the original spring embedder algorithm, which does not explicitly incorporate Hooke's law, Kamada and Kawai's algorithm moves vertices into new positions one at a time, so that the total energy of the system is reduced with the new configuration. It also introduces the concept of a desirable distance between vertices in the drawing: the distance between two vertices is proportional to the length of the shortest path between them. Kamada and Kawai avoid a second potential for repulsion by using springs of different length and strength between every pair of vertices. Their specific choice of springs is governed by the assumption that the ideal distance between two vertices is the length of a shortest path between them, multiplied by the ideal length of a single edge.

Finding a global minimum is difficult in a large search space. The strategy used in Kamada and Kawai's algorithm is to find a local minimum first. Nodes are moved into new positions if the movement leads to the fastest reduction of the total energy. The procedure is repeated until it converges – when the maximum improvement is less than a small fixed threshold. Instead of moving all vertices in the direction indicated by the force acting on them, their optimization reduces the inherent potential energy of the model by changing one vertex at a time.

The Kamada-Kawai algorithm achieves faster convergence and can be used to layout networks of all sizes. However, to obtain an aesthetically pleasing layout it sometimes becomes necessary to use the Fruchterman-Reingold algorithm after the Kamada-Kawai generates an approximate layout.

#### **2.4.1.7 Inverted Self Organising Map Layout**

A major drawback of the force-directed approach is the complexity of calculating the forces between all pairs of nodes. An alternative approach is the Inverted Self-Organising Map (ISOM) by Meyer (1998). Although not strictly a force-directed layout, the ISOM layout uses the idea of filling the space evenly with nodes and of causing connected nodes to attract each other. A form of competitive learning is used to update node positions by using a random input vector or stimulus. The position of the closest node to the stimulus is updated, along with its topological neighbours. By using a restricted update neighbourhood, only a constant number of nodes are updated at each iteration (Chan, et al. 2003). Although it is difficult to bound the required number of iterations, necessary for complexity analysis, this algorithm has been experimentally verified to have nearly linear time complexity (Meyer 1998, Au, et al. 2004).

The advantages of the ISOM layout method, which has an extremely simple implementation, are its adaptability to different types, shapes, or dimensions of layout areas and even to different metric spaces. On top of this it consumes only little computational resources which render it comparatively fast. Even though ISOM is impressively fast, it often yields poor layout results when used on its own. The resulting visual appearance for complex graphs is often not as symmetric as the results from Kamada-Kawai. The primary difficulty lies in the fact that it does not explicitly optimize the node overlap aesthetic, resulting in an unreadable layout for some graphs, particularly larger ones.

#### **2.4.1.8 Tree Layout**

Force-Directed Layouts can be a good choice when you're drawing general graphs and you don't have domain specific or any other topological information about them. Tree layouts however are a special case of graphs: this means that tree layouts have

more constraints and therefore there's more contextual information for drawing a tree than for drawing a general graph. Typically tree layout arranges the nodes, starting from a specified node(s), into a tree-like structure. Trees can be laid out nicely in a plane in polynomial time. The performance of the tree layout is proportional to number of nodes in the layout. They are easy to understand and nicely support abstraction and aggregation. Also, trees are easier to grasp for the human eye than "general graphs". There's that intrinsic notion of hierarchy that can be used to draw aesthetically pleasing graph layouts.

Consequently, additional aesthetics rules have also been formulated for them. For example, nodes with equal depth should be placed on a same horizontal line; distance between sibling nodes is usually fixed, etc. The Reingold and Tilford algorithm (Reingold and Tilford 1981) for trees is a good example of a layout algorithm achieving these aesthetics goals. Isomorphic subtrees are laid out in exactly the same way, and distance between nodes is a parameter of the algorithm. On the other hand, the more straightforward and naïve algorithm for displaying a tree, consisting of distributing the available horizontal space to subtrees according to their number of leaves, actually fails to achieve some of the aesthetic rules listed above. Here in Figure 11 is an example of the radial tree layout:

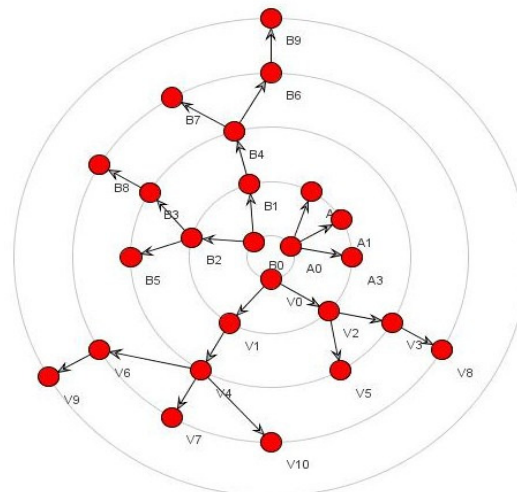


Figure 11: An example of the tree layout

In the implementation, this layout takes as input a Tree Graph to compute on internal structure. By combining this internal structure to a Walker algorithm, the algorithm places the graph nodes in hierarchically arranged layers where each parent node is centred on its sub-tree according to the hierarchical flow direction.



## Chapter 3

### Modelling Diffusion Processes in Social Networks

In this chapter, I present my work of approximating the diffusion processes in social networks. In section 3.1, I introduce the motivations of my work – why try to come up with a new method of describing the diffusion processes. After that, I describe my proposed method of approximating the diffusion processes in section 3.2. I then evaluate the accuracy of my proposed method by conducting experiments over real network data in section 3.3. In section 3.4, I investigate the influence measurement and ranking problem based on my proposed method. Finally, I discuss my work and give a conclusion in section 3.5.

#### 3.1 Motivations

Recall in section 2.2.1, I have categorized the three major challenges of modelling diffusion processes. One of them is how to capture social influence. In the diffusion processes, social influence determines to a large extent what we adopt and when we adopt it. Common approaches to address this problem include using a constant infection rate to encode neighbours' influence. As in the Independent Cascade Model (ICM) every time an individual contacts with an active neighbour, he has a constant rate of getting activated (Goldenberg, Libai and Muller 2001). However, many recent empirical studies (Leskovec, Adamic and Huberman 2007) have found that an individual's receptiveness to influence depends on past history of interactions with his neighbours as I have presented in section 2.2.2. Hence it is necessary to incorporate these empirical findings and propose adapted modelling frameworks.

On the other hand, as I have mentioned earlier, diffusion processes tend to start with a few early adopters of the behaviour and these early adopters may influence their friends, who may in turn influence their own friends and possibly lead to a cascade of influences (Kleinberg 2007). A natural question to ask is how many people will adopt the behaviour in the end. Formally, it is the problem of predicting the number of nodes reached by an initial target set (the early adopters of the behaviour) in the

diffusion processes. It is an open question to compute this quantity exactly by an efficient method. Typical method is to simulate the random diffusion process thousands of times to reach equilibrium and then average the result (Kempe, Kleinberg and Tardos 2003). However, according to Estevez, Vera and Saito this approach has a heavy computation load for large-scale network, which make it, may not be applicable in real world settings (Estevez, Vera and Saito 2007). Here I try to provide an alternative way of estimating the number of nodes reached by an initial target set in the end of the diffusion process.

### 3.2 Proposed Method

In this section, I describe my method of modelling the diffusion processes. It is argued that the process of new behaviours diffuse through social networks is very similar to the transmission of infected diseases (Leskovec, Adamic and Huberman 2007). Indeed, in the literature, it is common to describe the adoption process by using an infected disease spread analogy. Here I capture the diffusion process based on a non-linear dynamical system (NLDS) that accurately models virus propagations in epidemiology (Chakrabarti, et al. 2008).

Specifically in my method, given the initial active set and certain parameters we numerically calculate each node's probability to get activated. Following up to the Independent Cascade Model, we call nodes that adopt a behaviour is being active and inactive otherwise. Like many of the models I have discussed previously, there is explicit notion of dynamics or time in my model. It can tell us the probability that each node is activated at some point during the process and say nothing about the particular order in which the activation occur. The calculation of the probability is based on probability theory. For instance, the probability of node  $v$  is activated at current step is the probability of the event node  $v$  is not activated at previous step and  $v$  get infected from its neighbours happens at the same time. The calculation proceeds step by step until the increment of the sum of all probabilities is less than one, which means the number of nodes reached by the initial active set has been obtained. Formally, the method can be described as:

**Input:**  $G(V, E)$ ; Globe diffusion rate  $p_1, p_2$ ; Initial Active Set  $A$ .

**Output:**  $\sigma(A)$  - expected influence nodes count of  $A$ .



**Begin**

Start with time step  $t = 1$ .

While  $\sum P_{v,t+1} - \sum P_{v,t} \geq 1$  for all  $v \in V$  -  $\sigma(A)$  keeps increasing.

For all  $v \in V$  calculate the probability of  $v$  get activated at current time step  $P_{v,t}$ .

Get the sum of all probability values  $\sum P_{v,t}$ .

Go to next time step  $t + 1$ .

**End**

Where the social network is represented by a graph  $G(V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of edges. Let  $A$  denotes the set of initial active nodes at the beginning of the diffusion process, and it corresponds to the early adopters of a behaviour. Let  $\sigma(A)$  represent the number of nodes reached by the initial active set in the diffusion process. Let the probability that a node  $v$  is activated at time step  $t$  by  $P_{v,t}$ . Clearly, for all nodes in the initial active set  $P_{v,0}$  is 1 and  $P_{v,t}$  is 0 for the time step afterwards. For the rest of nodes in the network  $P_{v,0}$  is equal to 0. Where  $p_1, p_2$  denotes the globe diffusion rate.

As one can see from the pseudo code, the central of my method lies in the computing of each node's probability to become active each time step. Let's start with the compute of node  $v$ 's probability to become active at time step  $t$ . Considering node  $v$ 's neighbour node  $w$ , node  $v$  has no chances of getting infected from node  $w$  is either because node  $w$  is inactive at previous time step or node  $w$  is active but failed to infect node  $v$  through the link they are connected by with probability  $1 - \theta$ , where  $\theta$  be the infection catch rate on a link connected by an infected node. Therefore the probability that node  $v$  has no chances of getting infected from  $w$  is  $P_{w,t-1}(1 - \theta) + (1 - P_{w,t-1})$  which is  $1 - \theta * P_{w,t-1}$ . It is the probability of the event that node  $w$  is active at time step  $t - 1$  and succeeded in infecting node  $v$  through the link they are connected by has not happened. Assuming the behaviour that each neighbour attempting to infect node  $v$  is independent of each other, hence the probability of node  $v$  has a chance of getting infected from any of its neighbours is:

$$1 - \prod_{w \in N(v)} (1 - \theta * P_{w,t-1}) \quad (3.1)$$

Where  $N(v)$  denotes the set of neighbours of node  $v$ . Hence node  $v$  becomes active at time step  $t$  if node  $v$  is inactive at time step  $t-1$  and node  $v$  has a chance of getting infected from its neighbours:

$$P_{v,t} = (1 - P_{v,t-1}) * (1 - \prod_{w \in N(v)} (1 - \theta * P_{w,t-1})) \quad (3.2)$$

In epidemiology a susceptible individual can become infective on contact with another infective individual, and then heal herself with some probability to become susceptible again (Chakrabarti, et al. 2008). Here I only focus on the case where an active node can not be switched back to be inactive as this scenario is more common. Taking customer churn in business analytics as an example, companies typically consider churning customers that come back to the network as new customers.

Note  $\theta$  stands for the probability that a node get infected through a particular link with an activated neighbour. It is frequently assumed in epidemic models that individuals have equal probability of being infected every time they interact. According to Leskovec, Adamic and Huberman this may not be right (Leskovec, Adamic and Huberman 2007). Through observing the propagation of recommendations on a person-to-person recommendation network they found out that the probability of activation decreases with repeated interactions. For instance, if one of your friends recommended you to buy a product and you did not buy it, the next time your friends recommended it makes sense that you are less likely to buy it. This observation is somehow consistent with the decreasing cascade model (Kempe, Kleinberg and Tardos 2005), in which a node's propensity for being activated changes as a function of which of its neighbours have already attempted (and failed) to influence it.

Inspired by these studies, I attempt to encode the rule that the effectiveness of the influence through a particular link changes as the calculation unfolds and it depends on the past history of interactions. When a node first tries to influence its neighbour - its probability of being activated is not equal to 0 - we start to keep track of the number of trial times. Let  $k$  denotes the number of trial times. Let the probability that node  $v$  attempts to infect node  $w$  through their link by the first time be  $p_1$ , and by second and afterwards times be  $p_2$ , then

$$\theta = \begin{cases} p_1, k = 1 \\ (1 - p_1)(1 - p_2)^{k-2} p_2, \forall k \geq 2 \end{cases} \quad (3.3)$$

The reason I distinguish the first trial from the rest is motivated by a generalization of Independent Cascade Model (de Kerchove, et al. 2009) that considering different probabilities for being activated depending on the number of contacts with the information. Their results show that first and subsequent trials play different roles in the propagation process.

Given the network structure and specified value of  $p_1$  and  $p_2$  we can calculate the probability for each node to get activated at every time step with the specified initial activate set. The sum of all probabilities values will keep increasing as the calculation proceeds. When the increment is less than one the calculation will terminate, as that means the expected activated nodes count has been obtained. Meanwhile, when a node's probability of getting activated is less than the value of one divided by size of the network it will be considered as negligible, which means we no longer calculate its probability in the following time steps.

### 3.3 Experimental Validation

Having described my proposed method, I will focus on understanding its behaviour in practice. As I have mentioned earlier, an important quantity diffusion models need to predict is the number of nodes reached by an initial target set. Traditionally this is done by simulating the random diffusion process thousands of times to reach equilibrium and then average the result (Kempe, Kleinberg and Tardos 2003, Estevez, Vera and Saito 2007). In my method, I explicitly calculate this quantity. Hence here in order to validate my proposed method, we need to check whether it can give a good prediction.

Specifically, using two coauthorship network data I evaluate the accuracy of my proposed method of modelling diffusion processes by comparing its prediction performance against diffusion simulations. I examine the time evolution of the activated nodes count at both my method and the diffusion simulation. Further, I compare the activated nodes count at the end in both cases. Experiments show that

my proposed method yields very close results comparing to the diffusion simulations.

### **3.3.1 The Network Data**

In my experiment, two different size data sets of scientific coauthorship network were tested. It has been argued extensively that coauthorship networks capture many of the key features of social networks more generally (Kempe, Kleinberg and Tardos 2003). The first one is a coauthorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006. It contains 1589 nodes and 4331 edges. This data set will be referred as NS dataset in my experiments. The second one is a weighted coauthorship network between scientists posting preprints on the High-Energy Theory E-Print Archive. There are 8361 nodes and 15751 edges in the network. It consisted of 581 connected components, and the number of nodes in the largest connected component is 5835. It is a scale-free network with a power-law degree distribution. This data set will be referred as Hep-th dataset in my experiments. Both of the networks were obtained from Mark Newman's network data collection (<http://www-personal.umich.edu/~mejn/netdata/>).

### **3.3.2 The experiments**

I measure the diffusion process by examining the time evolution of the activated nodes count as the dynamic process unfolds. The diffusion simulation was used as a baseline to validate the accuracy of my method in modelling diffusion process. More specifically, I keep records of the overall activated nodes count in the network at each time step in both cases, then check whether they are close to each other approximately. In the calculation case, the overall activated nodes count refers to the integer value of the sum of probabilities of all nodes in the network at current time step. In the simulation case, at each random process the overall activated nodes count at each time step was stored. In Kempe, Kleinberg and Tardos (2003)'s work, the random process will reach equilibrium after 10k simulations. My result is consistent with that. Experiments indicate that the result of 10k times is comparable to 100k times or more. The overall activated nodes count at each step then estimated by get a mean value of the 10k simulation results.

Meanwhile I make sure the experiments are conducted with the same set of initial active nodes on a given network topology. Parameters  $p_1$  and  $p_2$  should be specified in advance like in the Independent Cascade Model. The value of  $p_1$  should be the same as the universal diffusion rate in Independent Cascade Model, with typically value 10%, and the value of  $p_2$  should be smaller than  $p_1$ . The simulation was conducted using the Independent Cascade Model. In the NS dataset, the universal diffusion rate for the ICM simulations is 10%, and the  $p_1, p_2$  value for my proposed estimation method is 10% and 7.8%, respectively. In the Hep-th dataset, the universal diffusion rate for the ICM simulations is 9%, and the  $p_1, p_2$  value for my proposed estimation method is 9% and 8%, respectively. The initial active set  $A$  was chosen randomly. With different initial active sets, the results were almost the same – two curves are similar two each other. Below in Figure 12 and 13 shows the time evolution of the activated nodes count in both simulation and calculation in the two datasets.

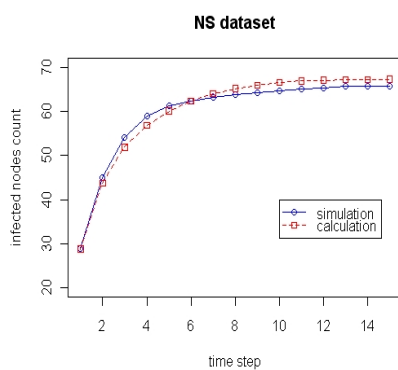


Figure 12: The plots show the time evolution of activated nodes count in the diffusion process with both simulation and calculation in the NS dataset.

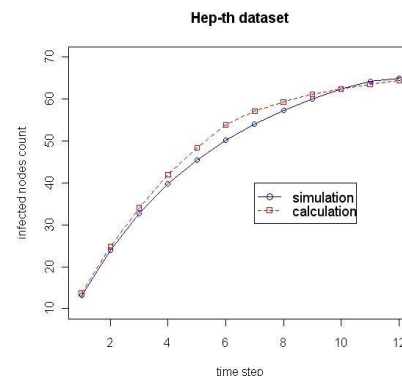


Figure 13: The plots show the time evolution of activated nodes count in the diffusion process with both simulation and calculation in the Hep-th dataset.

As we can see from the figures, in both datasets the two curves exhibit qualitatively similar shapes, dominated by a diminishing returns property in which the curve continues increasing, but more and more slowly and eventually flattened. As shown, my method nicely tracks the simulation results.

However, with regard to the computational complexity, the calculation did not outperform the simulation as I expected. Though it does not need to run the random diffusion process thousands of times, my proposed method still requires a heavy computational load. This probably attributes to the calculation of each node's probability to get activated at each time step, as it involves large amount of java floating point arithmetic, which can be time consuming.

### **3.4 Measuring Influence and Identifying Influencer**

As we stated previously, social network analysis is fast emerging as an important discipline for predicting and influence consumer behaviour. Let's consider the scenario when companies want to start a viral marketing campaign, which set of customers should they target as trendsetters in the first place? A simple answer is influential customers but the question is how we identify those influential customers. A fundamental question lies in the applications of social network analysis is to establish reasonable influence measures and further pinpoint influential customers.

Recall in section 2.1, I have introduced various network centrality measures. A major problem in social network analysis is to identify the important actors at both the individual and group levels of analysis. Who is the most important or central actor in this network? There are many answers to this question, depending on what we mean by important. For instance, degree is a measure in some sense of the popularity of a node. Closeness measures how close a node is to all other nodes in the graph. Betweenness is a measure of the extent to which a node lies on the paths between others.

In the customer network scenario, intuitively, high connectivity in the network could be a factor as the commonly used centrality based heuristics in the sociology literature (Freeman 1979). However, finding the customers that influence the buying decisions is not trivial. A customer who is not widely connected may in fact have high influence value if one of the neighbours is highly connected (Domingos and Richardson 2001). Customers influence does not end with the immediate neighbours. Those neighbours may in turn influence their own neighbours and possibly lead to a cascade of influence. This is closely related to the diffusion processes in social

networks. Clearly, the number of customers reached by that customer in the diffusion process could be an important influence indicator.

In this section, I look into the possibility of measuring nodes' influence value and identify influential nodes in social networks based on my proposed method. I use the number of nodes reached by an initial target set in the diffusion process as an indicator of the importance or influence of the initial target set. In the last section I have demonstrated the accuracy of method in modelling diffusion process. Hence I can use the sum of all probability values as an approximate estimation to the number of nodes reached in the diffusion process.

Specifically, the algorithm I used to calculate this quantity can be described as follows: The inputs needed are the network structure, proper parameters setting and initial target set. Starting with time step 1, calculate each node in the network's probability to get activated at current time step and sum up all the probability values, and then go to next time step. Repeat this process until the increment of all probability values is less than one. Here I calculate this quantity (For simplicity, I refer to it as influence value) for each node in the NS dataset, and further compare the influence value calculated with traditional centrality measures such as degree centrality. Below in Figure 14 shows the distribution of influence value for all nodes in the NS dataset; in Figure 15 compares the influence value and node degree centrality for all nodes in the NS dataset.

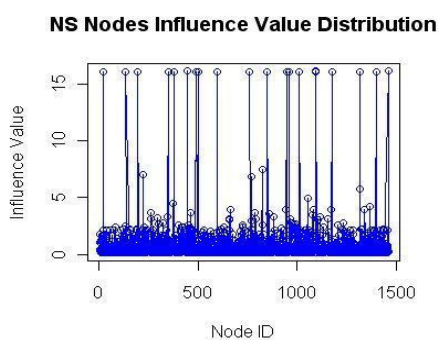


Figure 14: The plot shows the distribution of influence value for all nodes in the NS dataset.

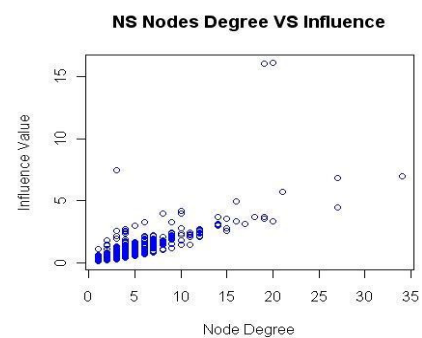


Figure 15: The plot compares the influence value and node degree for all nodes in the NS dataset.

As shown in Figure 14, in the NS dataset most of the nodes can only influence one or two nodes in the diffusion process, while only a few nodes can influence more

than two nodes. Intuitively, it seems beneficial to target those nodes that can influence more than two nodes in the diffusion process to spread viral marketing campaigns. As shown in Figure 15, node degree and influence value calculated are not fully correlated. For instance, there is one node with degree 4 and influence value around 8. High degree does not necessarily mean high influence value, and vice versa. This implies that the influence value metrics provides ranking methods that in general extract nontrivial nodes as influential nodes.

### **3.5 Discussion & Conclusion**

In this chapter, motivated by recent empirical findings, I extend an epidemic model that accurately models virus propagation to describe the diffusion process in social networks. With proper parameters setting on a given network topology, my proposed method can numerically calculate each node's probability to get infected when a set of nodes has been initially activated. By comparing its predicting performance with diffusion simulations, I validate the accuracy of my proposed method in capturing diffusion process. However, it does not outperform traditional diffusion simulations as I expected. When it comes to compute the number of nodes reached by set of initial active nodes, my model can give a suitable estimation to this quantity. Further, using this quantity as an influence measure, experiment results show that my proposed method provides ways of extracting nontrivial nodes as influential nodes.

The development of theoretical models for diffusion process still remains to be an open question. First of all, are these models discussed above correctly captured the way influence spreads through real network? All the models take a snapshot of the network, and then operate upon this fixed snapshot. No dynamic aspects or network evolution involves – it does not consider the network growths. Also all the models unfold in discrete time step with each node following certain probabilistic rule, and it uses this rule to incorporate information from its neighbour over time. Whereas the dependence of probability of adopting behaviours on number of friends adopted expressed in this way reflects an aggregate property of the full population, and does not imply anything about any particular individual's respond to their friends' behaviours (Kleinberg, 2007). Secondly, the way such dynamic process is affected by the network structure is still poorly understood (Kiss and Bichler, 2008). How



adoption probability depends on the structural properties of a node's network neighbours? What role does weak and strong ties play in the dynamic process? Is information propagates more quickly on a dense network?

According to Backstrom (2006) it has to date been easier to explore such models theoretically than to obtain reasonable estimates for them empirically on large-scale data. My future work directions include obtaining actual information diffusion data and observing how influence propagates in real network. Therefore we could develop ways to infer or estimate relevant model parameters with the historical diffusion data (Saito, Nakano and Kimura, 2008). With the support of the empirical findings we could make more general assumptions on how individuals respond to friends' influence, which leads to a closer integration of the theoretical models to the empirical results. Meanwhile, as my method has not outperformed traditional diffusion simulation as we expected, in the future work we need to improve the calculation process so that it can be fast.

## Chapter 4

### Empirically Measuring Diffusion Processes

In this chapter, I present my work on empirically measuring the diffusion processes. I obtain data from a telecommunication operator and construct customer call networks using the call records data. I then analyze the call networks and assess social influence in the bundle adoption process. Specifically, in section 4.1, I present the introduction and motivations of my work. I describe the dataset I used to build the call networks in section 4.2 and how I built them in section 4.3. In section 4.4, I present the analysis results of the call networks' structure properties. In section 4.5, using bundle adoption data I assess social influence in the customer call networks. Finally, the discussion of my work and conclusion are given in section 4.6.

#### 4.1 Introduction & Motivations

In the last chapter, I have presented my effort to model the diffusion process and in section 2.2, I have introduced various diffusion models. While many of the models address the question of how influence spreads in a network, they are based on assumed rather than measured influence effects (Leskovec, Adamic and Huberman 2007). The wide variety of rules theoretical diffusion models posed on individuals' behaviour, even if plausible, are often lacking empirical support (Cointet and Roth 2007). Furthermore, most of the diffusion models take a single snapshot of the evolving network and then build upon this static network topology. Even though there is time step in the simulation of the dynamic process, it has nothing to do the actual time of information spread. As such, it becomes unclear how accurately existing models render real-world diffusion phenomena. Cointet et al. (2007) suggests that future investigations of the diffusion mechanisms should begin with adequate empirical protocols; then propose adapted modelling frameworks.

On the other hand, as I stated previously, the emergence of mobile, email and online social networking have gradually transformed communication among people. Social interaction through these platforms leaves extensive digital traces by its very nature.

At unprecedented levels of scale and resolution we can now observe and quantify human communication patterns (Kleinberg 2008). Hence, it is necessary to obtain real world network data and conduct empirical analysis and further characterize social interactions.

The analysis of telephone call graphs is an exciting area of research, because phones are ubiquitous, they have become a strategic component for modern life and modern economies, and they are expected to become a key or even the principal conduit not just for voice calls, but for internet access and use in the future as well. Furthermore, they can provide detailed information on the spatio-temporal behaviour of users, especially on the social networks they build and maintain, as reflected by their phone calls. Indeed, several recent studies (Onnela, et al. 2007, Dasgupta, et al. 2008) have used call graph data to examine and characterize the social interaction of phone users, with a focus on understanding the structural properties of the graph, its evolution and the evolution of social groups, or the spread of new products and services.

In the following sections, I obtain four months call records data from a landline telecommunication operator. I consider the social networks induced by the calls of users in phone network. I present the procedures of how I build the call networks using the call detail records data. The call networks retrieved from the Call Detail Records can be used to examine and characterize the social interactions of the phone users. A social relationship between two subscribers, in this context, is based on the duration of the voice calls, call frequency that are exchanged during a certain period of time. I then present a detailed analysis of the call networks by examining its degree, connected component distribution. After that, I assess social influence in the phone users in the bundle adoption process.

## **4.2 Data Description**

The main information in respect to the social network analysis is certainly the data which establish the edges among the nodes. In the telecommunication scenario, this is the records which contain the calls among the customers, representing the way they relate to each other. This type of call in telecommunication is called CDR – Call Detail Record. CDR includes sufficient information to describe the important

characteristics of each call. At a minimum, each CDR will include the origin and destination phone numbers, the time and duration of the call. In a sense, call records data reflects the interaction patterns among telecommunication customers. Here I emphasize that my interest is in an aggregate statistical analysis and therefore, I do not study any particular customer's calling pattern. I analyze the CDRs and restrict my focus to the patterns of calls and networks formed out of these calls.

As I was involved in a project that collaborated with a landline telecommunication operator, I can get access to their call records data and analysed it for research purpose. Specifically, I consider four consecutive month (from October 2008 to January 2009) CDRs of 181, 890 sample residential customers of a landline telecom operator, which is around 20% of the whole residential customer base. In each month, there are around 10 million call records generated by the sample residential customers. These customers' calls are extracted from the corporate data warehouse. In my analysis, in order to maintain privacy and anonymity, data that could identify telecom customers is not utilized and I analyze the anonymized CDRs.

### **4.3 Build Call Networks**

I construct call graphs from the dataset described above. The nodes represent telecommunication customers, and edges indicate the calls exchanged between the customers. As my goal is to examine the social interactions among the sample residential customers, I only consider calls exchanged within the sample customers group. Specifically, during pre-processing, calls to service numbers such as the operator's customer service number and numbers similar to 1-800 numbers have been filtered out. Calls involve other operators such as calls to mobile or foreign countries have been excluded. As I have full access to the customers of the landline operator, but only partial access to the activity of other providers. Also calls that are not exchanged within the sample customers group have been filtered out.

Further, in order to translate the data into a network representation that captures the characteristics of the underlying communication network, I only consider calls that can represent social relationships. Particularly, very short duration calls (for instance less than 5 seconds) have been ignored as missed calls or wrong calls. These numbers and calls greatly skewed the call distribution may yield incorrect results. It

is possible that this may induce some false negatives, for instance some links corresponding to genuine social interaction may go undetected. Meanwhile, if A and B are two nodes in the call graph, then an edge exists between A and B if and only if A and B have called each other at least once during the time period of investigation. A single call between two individuals may not contain much information. Therefore only mutual calls of long duration can be served as an indication of social relationship.

Note that there can be multiple calls between two customers; still there is only a single edge between the corresponding two nodes. I represent multiple calls between any two nodes by a single edge, which can be associated with a weight. I aggregate the total number of calls and the total duration of the calls between two customers. Then the multiplicity of calls can be interpreted as weight of the edges. High weight value means long duration and high frequency calls which indicate strong social ties.

In the datasets, take the call records in October 2008 as an example, the raw CDRs data contains 10,476,250 records. Eliminating service number calls and foreign countries calls and etc reduced it to 6,254,577 records. Further, considering calls exchanged only within the sample customers group reduced it to 485,783 records. By aggregating calls exchanged between two customers reduced it to 122,885 records. In the end, I get a call graph containing 85,421 nodes and 122,885 edges. In fact, in the original 181,890 sample residential customers, 26% of the customers do not generating calls during the examining period. Considering the customers' calls to the mobiles or to other customers not within the sample group, it is reasonable that in the end there are only 85,421 nodes in the call graph.

#### **4.4 Basic Network Characteristics**

Using the four consecutive month CDRs of the sample customers described above, I build four call graphs in a monthly basis. Table 5 displays the summary network statistics for the sample customers' four consecutive month call networks. In the table, LWCC stands for Largest Weakly Connected Component which means the size of the largest weakly connected component in the network.

Table 5: Summary network statistics for the sample customers' four month call networks.

Month	Nodes	Edges	Density	LWCC
October 2008	85,421	122,885	0.0000168	63,145
November 2008	84,868	121,594	0.0000169	62,000
December 2008	89,885	136,862	0.0000169	70,577
January 2009	87,173	127,590	0.0000168	64,793

From Table 5, we can see that despite minor changes on the sizes of nodes and edges of the four networks, the density and ratio of nodes of the largest connected component remain almost the same. The fluctuation of size of nodes and edges in the networks can be explained as some customers who have some usage in a particular month and do not present the same usage for the subsequently months. Customers do not necessarily have the same communication patterns every month. Roughly, I can conclude that the sample customers' call network is relatively stable during the examining four month period. Due to the sizes of the networks, it is challenge to validate this observation accurately. Based on this in the following investigation, I focus a particular month let's say October 2008 to study the structure properties of the call graph.

#### 4.4.1 Degree Distribution

What are the structural properties of the call graph? I begin by examining the degree distribution of the nodes. Figure 16 and 17 give sample customers' October 2008's node in-degree and out-degree distribution. Observing their in-degree and out-degree distributions, the call graph topology is found to be characterized by presence of a highly heterogeneous topology, with degree distributions characterized by wide variability and heavy tails. In general, the degree distributions are skewed with a fat tail, suggesting that only a few customers are related to large group of people while most of them call a relatively smaller number of people. This call graph characteristics is consistent with the power-law degree distribution exhibits in many graphs and social networks, which is a typical signature of scale-free networks.

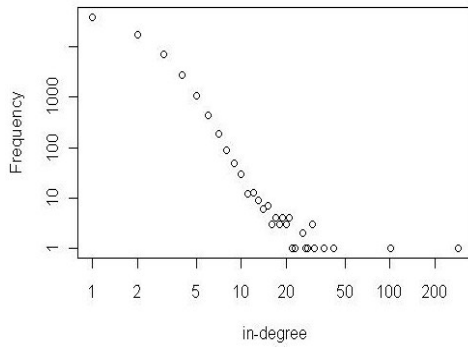


Figure 16: The in-degree distribution of the sample customers' October 2008 call graph.

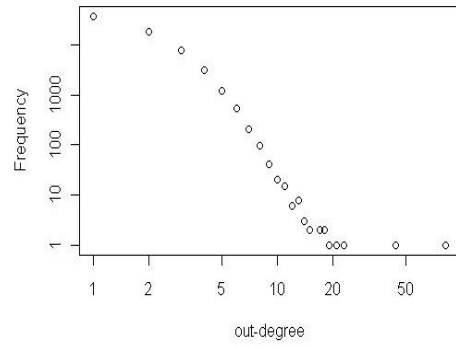


Figure 17: The out-degree distribution of the sample customers' October 2008 call graph.

#### 4.4.2 Connected Component Analysis

In a graph, it is said that a subset of the graph is a connected component if it is possible to go from each node in that subset to another node in the same subset by following links (in any direction). The subset is called a strongly connected component if this is possible by respecting the direction of the links otherwise is called a weakly connected component. I study the distribution of the sizes of the weakly connected components (WCCs) in the call graph. The distribution of the sizes of the components is presented in Table 6. From this table, we can see it contains a giant connected component, which indicates that there may be a core social group forming in the telecommunication customers' network. As the power law degree distribution displayed above, this is also a typical signature of scale-free network.

Table 6: Size of the weakly connected components in the October 2008 call graph.

Component size	Number of components
2	5461
3	1430
4	534
5	258
63145	(Giant component)1

Next I plot the sizes of the connected components in the call graph. As seen from Figure 18, a power-law distribution of the sizes of the connected components is observed.

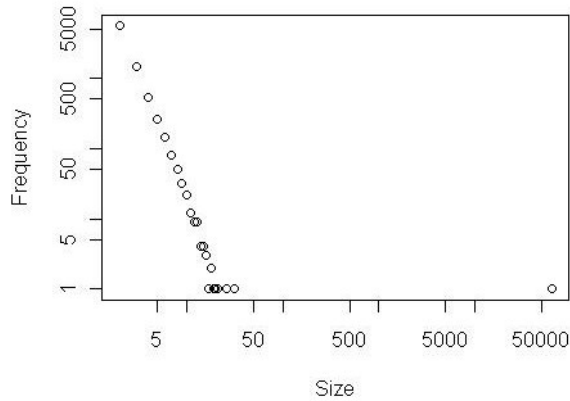


Figure 18: Distribution of the sizes of the connected components in the October 2008 call graph.

### 4.4.3 Persistence Analysis

As customers may have different usage pattern in different months, some nodes may appear in one month's call network and disappear in another. Firstly I investigate the persistence of the nodes. For each month, we say a node which is in that month's call network will have a persistence value one and otherwise value zero. As there are four months, for all the nodes their overall persistence value will be in the form of 1000, 0101, 1111 and etc. I view this in the binary format and convert it to decimal. Hence, for the four months each node will have an overall persistence value ranging from 1 to 15. Below Figure 19 shows the distribution of the nodes' overall persistence value. In particular, with value 15 means that node is in all the four month's network. Among the overall 112, 481 nodes, 61706 nodes (54.9%) are in all the fourth month's network, which somehow confirms the observation that sample customers' call networks are relatively stable during the examining four month period.

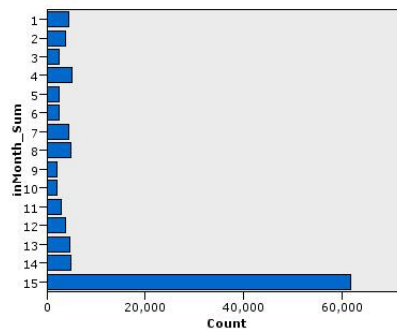


Figure 19: Distribution of nodes' overall persistence value.



Next, I investigate the persistence of the ties. I based my analysis on a link persistence measure introduced in a study that explores the correlation between the structure of a mobile phone network and the persistence of its links (Hidalgo and Rodriguez-Sickert 2008). Specifically, they measure the stability of ties across time as the number of panels in which a link is observed, over the total number of panels, which can be expressed as:

$$P_{ij} = \frac{\sum_T A_{ij}(T)}{M} \quad (4.1)$$

Where  $A_{ij}(T)$  is 1 nodes  $i$  and  $j$  communicated on panel  $T$  and 0 otherwise, where  $M$  is the total number of panels. Persistence is the probability of observing a tie when observing a panel of network data. In my case,  $M$  will be four and the tie persistence value will be  $1/4$ ,  $2/4$ ,  $3/4$  and 1. Below in Figure 20 shows the distribution of the tie persistence value. Among the 268, 481 ties, 55.4% has persistence value 0.25; 16.6% has persistence value 0.5; 10.7% has persistence value 0.75; 17.3% has persistence value 1.

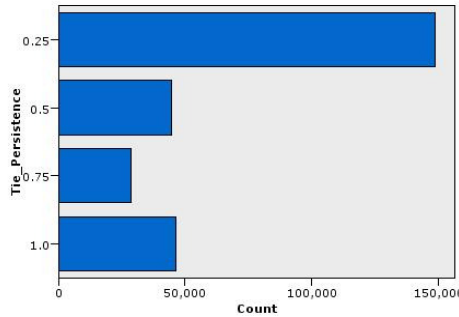


Figure 20: Distribution of the tie persistence value.

## 4.5 Measuring the diffusion

It has been a challenge to build reasonable models for the patterns of communication within a social network: it is difficult to obtain data on social network structure at a large scale, and more difficult still to obtain complete data on the dynamics of a network's communication events over time (Kossinets, Kleinberg and Watts 2008). Here I focus on communication events that ripple through many nodes over short-scales following a particular event or activity. Basically, the analysis is performed

considering a particular business events happened in the past, and the evaluation of the afterwards events in a specific timeframe. For instance, in relation to the bundle adoption event, a particular month is taking into consideration and then, all customers who adopt the bundle in that month are included in the dataset of starters or triggers customers. Therefore, the subsequent events are analyzed considering the related customers, which means, the customers who have some relationship with the starter or the trigger ones. Typically, I am interested in the related customers' behaviour with regard to the event. For example, the evaluations of how many customers from the related ones have followed the starters or the triggers in the same event.

I obtain data related to customer broadband bundle adoption from the corporate data warehouse. The operator's data warehouse has detail records of the date the customers join the broadband and for those customer quit the broadband the date they cease. Specifically, in the 181, 890 sample customers during the four month examining period 97, 212 around 53% customers were once with the broadband. Among them 26, 847 around 28% customers has ceased till the end of the examining period.

#### **4.5.1 Correlation Analysis**

As I have stated earlier, an underlying premise many diffusion studies build upon is assuming the existence of social influence – people are influenced by their friends in the social network when making adopting decision. To understand the characteristics of bundle adoption behaviour and relate it to a diffusion process, I first need to find out whether there is any evidence of social influence in affecting customers' adopting behaviour. I use the correlation analysis to assess the existence of social influence in customers' call networks. The correlation analysis is basically the evaluation of how many customers, from the related ones, have followed the starters or the triggers in the same event.

More specifically, the correlation assessment is performed evaluating how many related customers have followed the starters in the same event in the subsequently months. The same analysis is performed for the random subset of customers, evaluating how many related customers have followed them in the same event in the

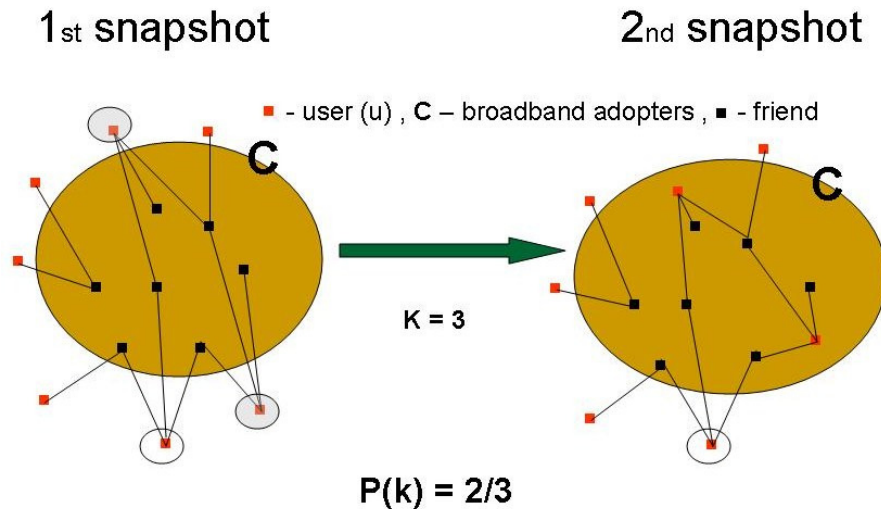
subsequently months. By comparing the related customers of the triggers and the random ones with regard to the same event, I would like to check whether friends of the triggers are more likely to participate in the same event.

I consider the set of customers who adopt the broadband during October 2008 and also in that month's call network as triggers. There are 2516 customers and they are related to 4804 customers in October 2008's call network. Among the related 4084 customers, 3541 of them (73.7%) are once with the broadband. And 444 of them (12.5%) have actually adopted the broadband in the subsequently months. As a comparison group, I select random 4804 customers from October 2008's call network. Among them, only 1483 customers (30.9%) are once with the broadband. And 154 of them (10.4%) have actually adopted the broadband in the subsequently months. From the ratios, we can easily see that friends of the triggers have a higher likelihood of adopting the broadband comparing to random ones. This result somehow indicates the existence of social influence with regard to the broadband adoption in the customer call networks.

#### **4.5.2 Dependence on Number of Friends Analysis**

Having confirmed the existence of social influence in customers' call networks with regard to bundle adoption in the last section, in this section I would like to examine the characteristics of the influence. If we view the diffusion process as a cascade of social influence, a natural starting point is to understand the local mechanism of the influence. In particular, I would like to investigate how does one's probability of adopting a new behaviour depends on his friends who have already engaging in the behaviour. Start with the easiest, the number of friends who have adopted the behaviour.

Recall in section 2.2, I have introduced various empirical studies that measure to what extent social influence affects the adopting decision. Here I adapt the dependence on number of friends analysis from a study on the group formation in the online communities (Backstrom, et al. 2006). Figure 21 below illustrates the process of the dependence on number of friends analysis.



Probability  $P(k)$  of adopting broadband = fraction of triples  $(u, C, k)$

Figure 21: Dependence on number of friends analysis.

In my context, the events are joining broadband or ceasing broadband. Firstly, I consider the join broadband event. Formally, Let  $p(k)$  denotes the probability of a customer adopt the broadband with  $k$  friends who have already adopted the broadband. Let  $C$  represents the set of customers who have adopted the broadband. Set  $U$  are the customers of interest; a customer  $u \in U$  is a customer who at the time of the first snapshot, did not belong to set  $C$ , and customer  $u$  had  $k$  friends in set  $C$  at that time.  $p(k)$  is then the fraction of such triples  $(u, C, k)$  for a given  $k$  such that  $u$  belonged to  $C$  at the time of the second snapshot. In figure 6, at the time of the first snapshot, among 3 customers in set  $U$  with given  $k$  is 3, 2 customers are in set  $C$  at the time of the second snapshot; Hence  $p(k)$  is  $2/3$ .

Back to the four month snapshots of the sample customer's call networks, using the November 2008's call networks, consider the customer join broadband event in December 2008. In this scenario, set  $U$  are the total 181, 890 sample customers minus the customers who once join broadband before December 2008's snapshot and the size of set  $U$  is 110, 568. Set  $C$  are customers who once join and still with the broadband before December 2008's snapshot and there are 65,530 customers in set  $C$ . Here I call customers in set  $U$  who actually join broadband in December 2008 as set  $W$ . In this case, there are 5,015 customers in set  $W$ . In set  $U$  50, 400 customers havenon-zero  $k$  value. In set  $W$  1,094 customers havenon-zero  $k$  value. These indicate the number of friends the broadband related customers have in the call networks. Some of the broadband related customers or their friends may not

in the call networks. Below Figure 22 and 23 shows the distribution of  $k$  value in set  $U$  and set  $W$ .

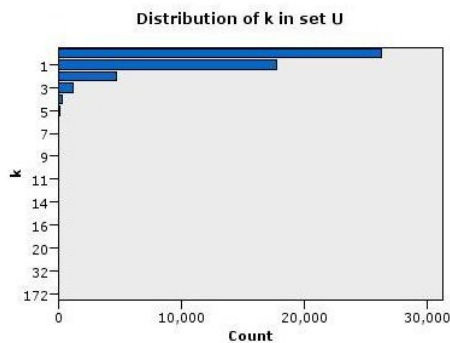


Figure 22: Distribution of  $k$  value in set  $U$  in join broadband event.

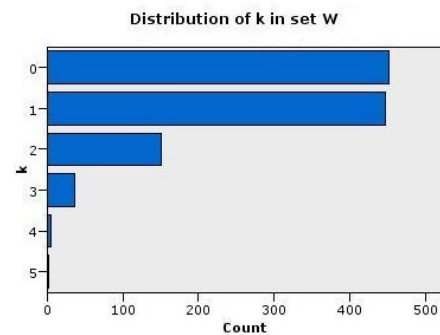


Figure 23: Distribution of  $k$  value in set  $W$  in join broadband event.

Below Figure 24 shows the distribution of the function  $p(k)$ . X-axis is the value of the  $k$  and Y-axis is the value of the  $p(k)$ .

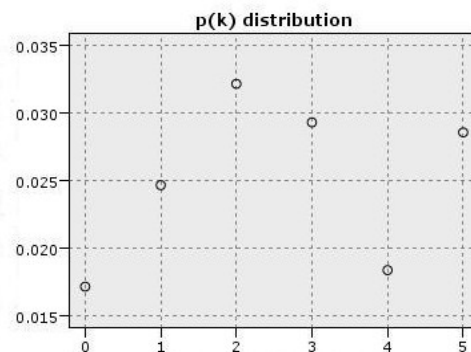


Figure 24: Distribution of the function  $p(k)$ .

Recall in section 2.2, as it has been observed in many studies (Backstrom, et al. 2006, Leskovec, Adamic and Huberman 2007), the diffusion curves should keep increasing – as more and more your friends adopting the behaviour, the probability of you also adopt the behaviour will increase. However, this is not the situation here –the curve displays a different shape. As we can see from the Figure 24, when  $k$  is 3 and 4,  $p(k)$  is decreasing instead of increasing. Naturally I expect the probability value should keep increase, however in fact it does not.

## 4.6 Summary and Discussion

In this chapter, I obtain four months call records data from a landline telecommunication operator. I consider the social networks induced from the calls of users in the phone network. The call networks retrieved from the Call Detail Records are used to examine and characterize the social interactions of the phone users. By analyzing the basic network characteristics such as degree, connected components distribution, nodes and ties persistence of the four month's call networks, I have found that the call networks are relatively stable during the examination period.

Using the bundle adoption data, I assess social influence in the call networks. By comparing the friends of the triggers (early adopters of the bundle) and random ones with regard to the bundle adoption event, we have found that friends of the triggers are more likely to also adopt the bundle. This result somehow confirms the existence of social influence in the call networks. Further, I adapt the dependence on number of friends analysis to investigate how does one's probability of adopting depending on number of friends who have already adopted. However, the result is not as expected that the probability keeps increasing as more friends adopted, as it has been observed in many other scenarios.

This result makes us to question how the call networks capture the underlying social interactions of the phone users. There may be several reasons: (1) when I construct the call networks, I only examine a sample of 20% of the whole residential customer base and I only consider calls within the sample customer groups. A customer's complete social interactions may not be well captured. (2) As I use landline phone records to approximate the social interactions of the phone users instead of mobile records, it is generally believe that mobile calls are more likely to represent social relationships. (3) As I only consider the overall statistical analysis result, in fact a customer's adopting decision may also be affected by other factors such as the price of the bundle and the quality of the service. At this stage I don't know is the social influence or other factors that influences customers adopting behaviour.

In the future work, I would like to obtain more detail data that allow us to better capture the underlying social interactions within the phone users. Meanwhile, I would like to investigate other business events such as customer churn to assess

social influence and measure the diffusion process, to see if the same kind of characteristics is exhibited in other scenarios.

## Chapter 5

### Social Network Analysis & Visualization Tool

In this chapter, I present my work on the prototypical tool for social network data manipulation, analysis and visualization. Firstly, I introduce the motivations behind my work and give an overview of the tool. Then I present the set of network analysis measurements and network visualization techniques I have implemented in my tool in detail. Specifically, in section 5.1, I describe the motivations of my work – why I try to come up with a new tool. In section 5.2, I give an overview of the prototypical tool describing its main features and implementation details. After that, I highlight the analysis features I have implemented in section 5.3. I present my efforts in network visualizations in detail in section 5.4. Finally, I conclude with the discussion of challenges and future directions for my work in section 5.5.

#### 5.1 Introduction & Motivations

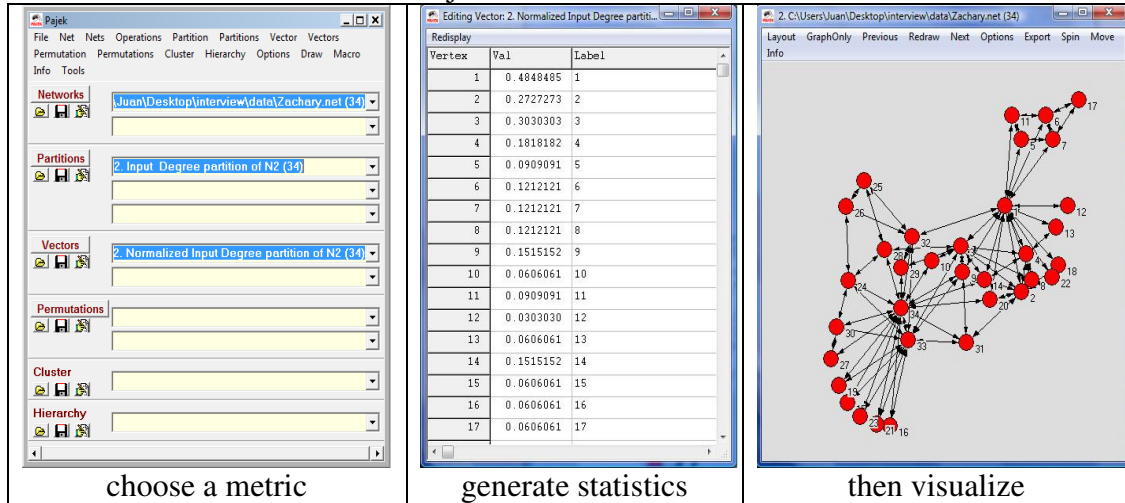
Network structures are important in many disciplines and professions. Interest in these structures is growing more common as the world of social networks and computer-mediated social content becomes more main stream. Recall in section 2.3, I have introduced various network analysis and visualization software tools. Researchers have created tools from set of network analysis components not limited to R and the SNA library, JUNG, Guess and Prefuse. There are other network analysis tool like Pajek, UCINet, and NetDraw that provide graphical interfaces, rich libraries of measures, and do not require coding or command line execution of features. So why create another network analysis tool?

The maturity of social network analysis tools has not advanced as fast as the popularity of social network analysis (Smith, et al. 2009). There are large amount of data revealing communication patterns among people, but the analysis tools for understanding the data have lagged. Many of the network analysis tools are designed for expert practitioner, have complex data handling, and inflexible graphing and visualization features that inhibit wider adoption. While some of the tools are simple



they still represent a significant overhead for domain experts who need to acquire technical skills and experience in order to explore data in their specific field. Below in Table 7 is a typical analysis and visualization routine with traditional social network analysis tool Pajek:

Table 7: Traditional SNA tool Pajek.



As we can see from the figures, the main user interface of Pajek contains the set of Networks, Partitions, Vectors and etc currently under investigation. Clearly, to use the tool, one needs to first understand what Networks, Partitions and Vectors are and prepare data in the specific formats. In fact, different tools may have different settings and it might take a while for a novice user to get familiar with the environment. Besides, the visualization component is separate from the main user interface. One has to perform certain operations to get a visual representation of the network under investigation. I believe a design that allows users to first grasp a visual impression on the network should be helpful. Hence, it is necessary to come up with a simple tool that can allow users with little knowledge on social network analysis to perform analysis and visualization tasks.

On the other hand, instead of focusing on coming up with extensive and complicated analysis and visualization features, I believe allowing users to interactively explore and manipulate network data should be a high priority. Adam Peter and Ben Shneiderman (2006) argue that interactive exploration of networks is currently challenging because: (1) it is difficult to find patterns and comprehend the structure of networks with nodes and links, and (2) current systems are often a medley of statistical methods and overwhelming visual output which leaves many analysts

uncertain about how to explore in an orderly manner. They present the SocialAction(Perer and Shneiderman 2006) tool which provides real-time exploration, filtering and clustering functions by integrating statistics and visualization. Using case studies they demonstrate that the tight integration of statistics and visualizations improves exploratory data analysis and dramatically speed insight development (Perer and Shneiderman 2008a).

I also believe allowing users to effectively develop insights on the networks under investigation should be the ultimate design goal for social network analysis and visualization tools. Instead of focusing on coming up with complicated analysis and visualization features, my strategy in designing the tool is more focused on improving interactive manipulation and exploration of the network data. Overall, my goal is to create an extendible network analysis tool that encourages interactive overview, discovery and exploration through “direct” data manipulation, graphing and visualization.

## **5.2 Overview of the Tool**

In this section, I give an overview of my prototypical tool for social network data analysis, visualization and exploration. It builds upon JUNG (Java Universal Network/Graph Framework) with rich support for dynamically manipulate and edit network data. First, I present its implementation details. Then I introduce some of the main features and techniques I have implemented in my tool to allow users to interactively manipulate and explore network data.

### **5.2.1 Implementation Detail**

My prototype tool is implemented in Java and it integrates several open-source libraries to take advantage of their contributions. JUNG provides my underlying node-link data structures, as well as an implementation of most of the network analysis and visualization algorithms. JUNG (<http://jung.sourceforge.net/>) is a java library that provides a common and extendible language for the modelling, analysis, and visualization of data that can be represented as a graph or network. It provides a common framework for graph/network analysis and visualization which allow users to build their own social network analysis tools. Jdesktop’sswingx library has been

used in the user interface development. Below in Figure 25 is a screenshot of the main user interface of my prototypical tool:

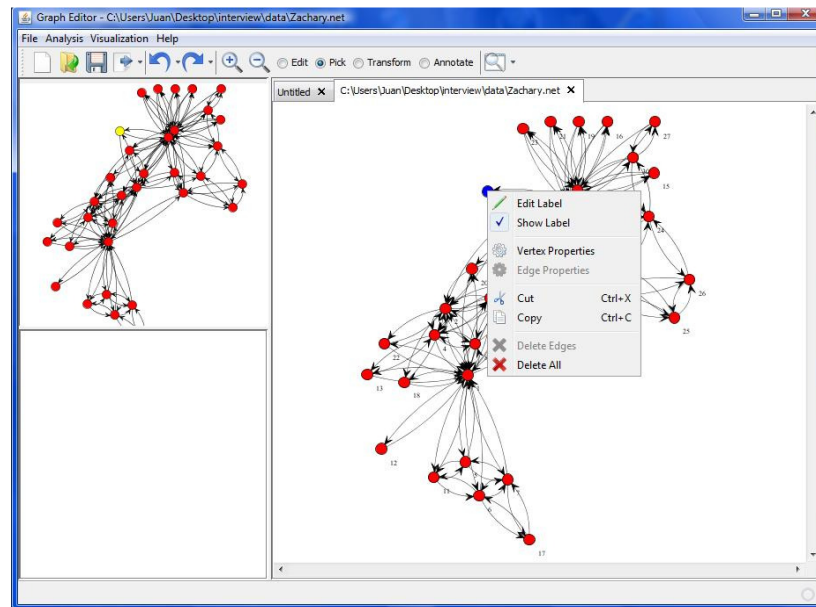


Figure 25: Main user interface of the prototypical tool.

As shown in the figure, the overall appearance and many concepts of user interaction are similar to many other social network analysis and visualization tools. The main window contains the standard elements of typical graphical user interfaces: a menu bar and a toolbar on top. Adhering to the common practice, the menu bar provides access to all functions, grouped into the self-explanatory menus File, Analysis, Visualization and Help. In the toolbar, the most frequently used operations are duplicated for user convenience. Most of the window's area is reserved for the main views of the open networks. Each network is displayed in full detail in its own view in a separate tab in order to allow users to work on related data simultaneously and to compare results of different networks or even different analyses of the same network. The main view depicts the network in full detail and allows interactive changes to the network structure and the graphical properties of the nodes and edges.

## 5.2.2 Main Features

In this section, I first describe a sequence of operations from data import to computation of network statistics and refinement of network visualization. Then I highlight some of the network data manipulation and visualization features.

Starting with importing the data, the users first need to load the data in pre-defined network format. In current implementation, data can be imported from Pajek files, GraphML files and comma separated value (CSV) files. When the network data was loaded, a visual representation of the network with Fruchterman-Reingold graph layout will be displayed in the main view. Sometimes when the network is too large for the graph layout engine to display, an error will report. After successfully loading the data, the users now can perform various analysis and exploration tasks on the network. Specifically, these may include generate network statistics, add or delete nodes or edges, customize the visualization and etc. The modified network can be saved in the Pajek, GraphML and CSV files. The generated network statistics can also be saved in CSV files. The visual image of the network visualization can be saved in typical image format such as JPEG, GIF, PNG and BMP.

Besides, in order to allow users to gain better control over the exploration and visualization of the networks, there are four mouse modes of operation of a view: Edit, Pick, Transform and Annotate. In the Edit mode, the main mouse operation is graph elements creation and users can add, delete nodes or edges, and modify their properties by simple mouse clicks. In the Pick mode, users can select the specific graph elements of interest and these may include nodes, edges and subgraphs. Note in order to edit the network structure, users must first go to the Pick mode to select the graph elements of interest. In the Transform mode, the main operations are the moving of the drawing area and zooming in and out. In this mode, neither the structure nor the layout of the graph can be changed. Throughout the process of exploring the network, users may come across important discoveries. In the Annotate mode, the tool features a light-weight solution for users to annotate these insights quickly. Often, annotations are textual comments such as indications of the insights. Note also the structure and the layout of the graph can not be changed in this mode.

### **5.2.2.1 Network Data Manipulation**

Although both statistical methods and visualizations have been used extensively by network analysts, exploratory data analysis remains a challenge (Perer and Shneiderman 2008a). In order to allow users to gain a full control on the networks under examination, I have a focus on allowing users to dynamically manipulate and edit the network data:

- Supports the selection of nodes, edges and different actions can be taken on selected graph elements. In the Pick mode, users can arbitrarily select certain subset of nodes or edges of interest. Operations such as change properties add or delete then can be taken on the selected graph elements.
- Users can dynamically add or delete graph elements (nodes, edges, subgraphs). Unlike many other social network analysis tools, the network being examined can not be modified. In my tool, users can change the network structure, add or delete nodes or edges. These actions are also undoable. The modified network can be saved in the standard network formats.
- Supports the Cut- Copy- Paste (transfer) actions on graph elements. Besides add or delete, transfer actions like cut, copy and paste are also supported on the selected graph elements. For instance, users can copy set of nodes and edges from this network paste them to the other.
- Users can easily edit attributes (shapes, colours, labels, positions etc.) of nodes and edges. Once certain graph elements have been selected, users can specify various visualization properties of nodes and edges. Below in Figure 26 is the screenshot of the user interface for setting node visualization properties. These properties comprise the shape, the colour, the size, the coordinates of nodes, and the width of edges and so on. For examples, the shape of the nodes can be square, oval, triangle and so on. The shape of the edges can be straight lines, dash lines, curve and so on. Additionally, there are extensive options to change the appearance like the font, positions of the labels of the nodes and edges. Below in Figure 27 is a sample network with different node visualization properties.

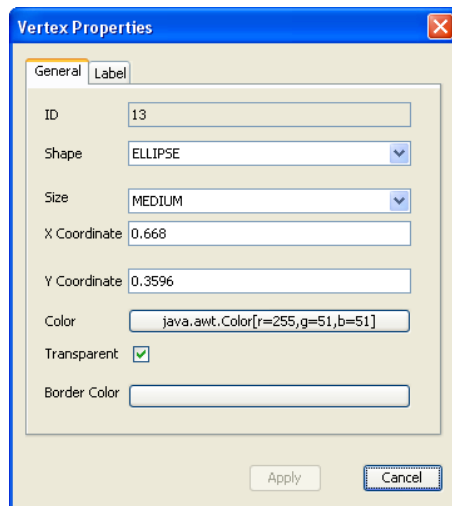


Figure 26: User interface for setting node visualization properties.

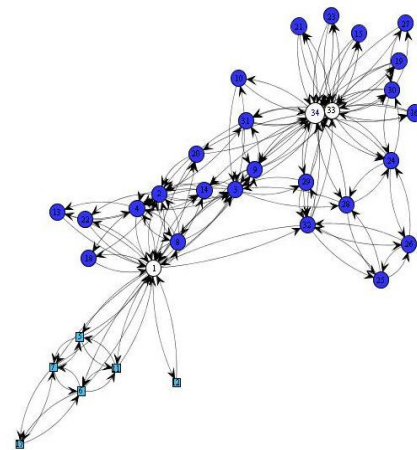


Figure 27: Sample network with different node visualization properties.

### 5.2.2.2 Visualization Features

Existing network visualization often seems impressive because of the colourful display of nodes richly connected with links. These visually engaging images enable users to estimate the network size while revealing important clusters. However, in most examples, the overlapped nodes prevent users from estimating cluster size and the crossed links make it possible to follow connections, count node in-degree, or carry out other tasks. Regarding to the network visualization, some efforts have been made to improve the interactively exploration and visualization of the network data:

- Rich support for manipulating and controlling the network visualization. In the Transform mode, the users can move the drawing area and zooming in and out the visualization. In the Annotate mode, the users can annotate the network visualization.
- Supports multiple classic graph layout algorithms such as Kamada-Kawai layout, Fruchterman-Reingold layout and Tree layout. Recall I have introduced in section 2.4, graph layout try to position the nodes of that graph using a layout algorithm that attempts to fulfil certain aesthetic requirements. In my tool, I have implemented several classic graph layout algorithms that are common in practice.
- Supports the satellite view in which the full graph is always visible and all mouse actions affect the graph in the master view. Visualizations of network

data are often cluttered, suffer from occlusion and illegible labels, and are difficult to explore. Zooming into sub-sections of the network can solve the problem somehow but may force users to lose the global structure. The satellite view can allow viewers to grasp the overall network structure when the displaying area is very large. Below In Figure 28 is the screenshot of a network visualization with a satellite view on the left upside pane. From the figure we can see in the satellite view the full graph is always visible.

- Supports the transform of network visualization via a “lens” effect. The basic idea of the lens effect is to use a tool like a magnifying glass to support enlarging the region of interest directly in the visualization. The area under the magnifying glass is processed using techniques such as fisheye projection, and the result is displayed differently than the remaining area. Overall, it shows a modified view of the selected region, while the rest of the visualization remains unaffected. Below in Figure 29 is the screenshot of a network visualization with a lens effect.

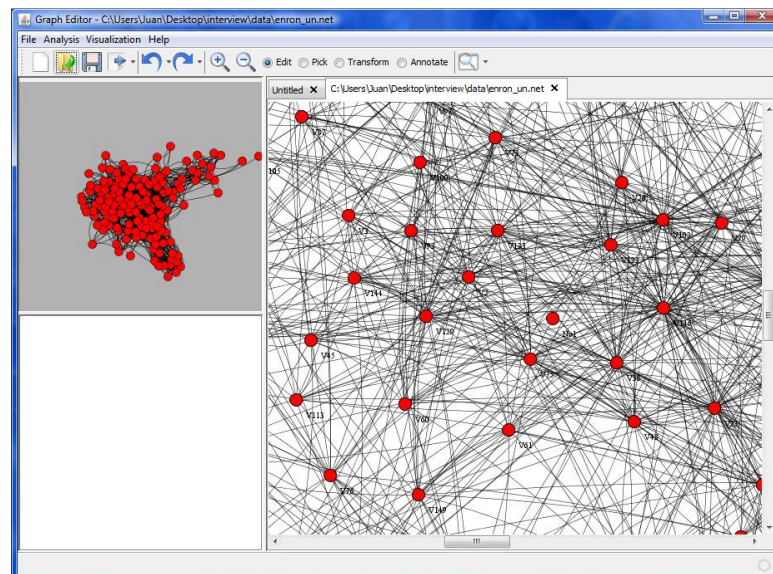


Figure 28: Screenshot of a network visualization with satellite view.

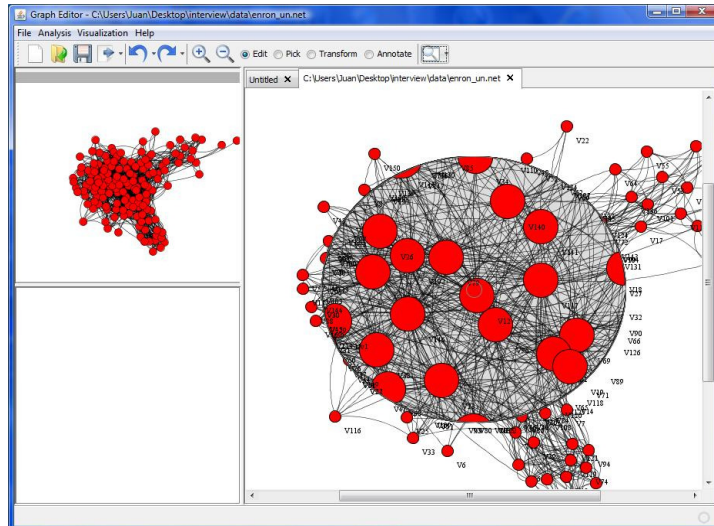


Figure 29: Screenshot of a network visualization with lens effect.

### 5.3 Analysis Features

In the last section, I have presented an overview of my prototypical tool. In the following sections, I concentrate on introducing its analysis and visualization features in detail.

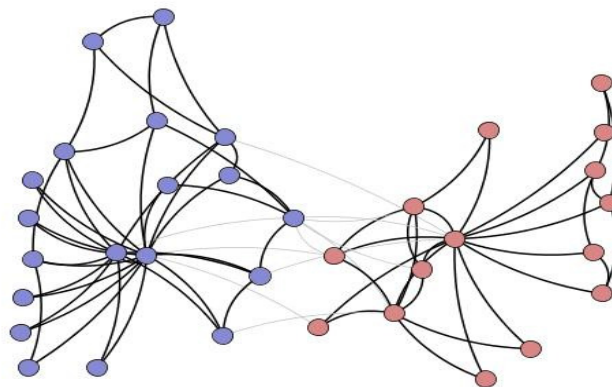


Figure 30: The edge betweenness community detection algorithm in my implementation.

Recall in section 2.1, I have identified two major goals of social network analysis: uncover notable nodes and discover cohesive subgroups. In the development of the analysis component of my tool, I try to follow these principles. In the literature, there are many statistical algorithms which reveal nodes that occupy key social positions and form cohesive social groups. Selectively, I have implemented some of the network measures that are most common in practice. These include various network centrality measures and some of the network community detection algorithms. Herein Figure 30 is an example of the implementation of the edge



betweennesscommunity detection algorithm in my tool. As most of these statistical algorithms have already been implemented in the open source library JUNG, all I need to do is to call the specific functions and add proper user interfaces.

On the other hand, social network analysis is inherently complex since analysts must understand every node's attributes as well as relationships between nodes. Despite many of the structuremeasures proposed to assess the networks, it is difficult to find outliers and patterns in strictly quantitative output. Often in these situations, network visualizations are often hard to interpret because overlapping nodes and tangled edges.

As I have introduced earlier, a tight integration of analysis and visualization could dramatically speed insight development (Perer and Shneiderman 2008a). The combination of visual data representation and statistical analysis support additional non-trivial observations about the data. An adequate visualization should highlight the graph structure and the result of the analysis measure simultaneously. With regard to the analysis component of my tool, I also try to follow this guideline. I try to visualize some of the analysis result. Here I use two examples to demonstrate the idea. The first one is to use the size of node to depict the degree centrality score of nodes as shown in Figure 31. The other one by assigning different nodes colour and location, it can differentiate the group each node belongs to as shown in Figure 32.

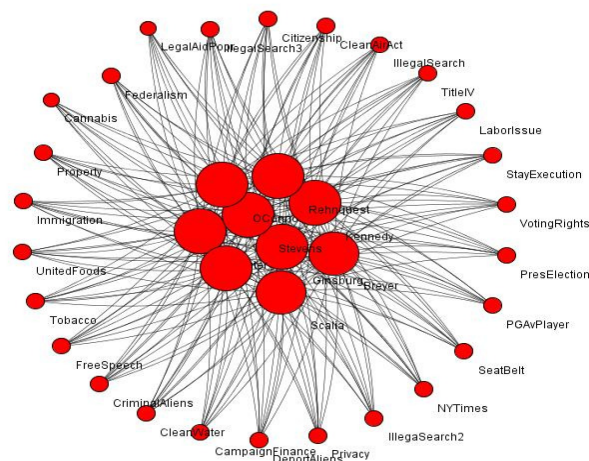


Figure 31: A network visualization with size of the node scaled to node degree.

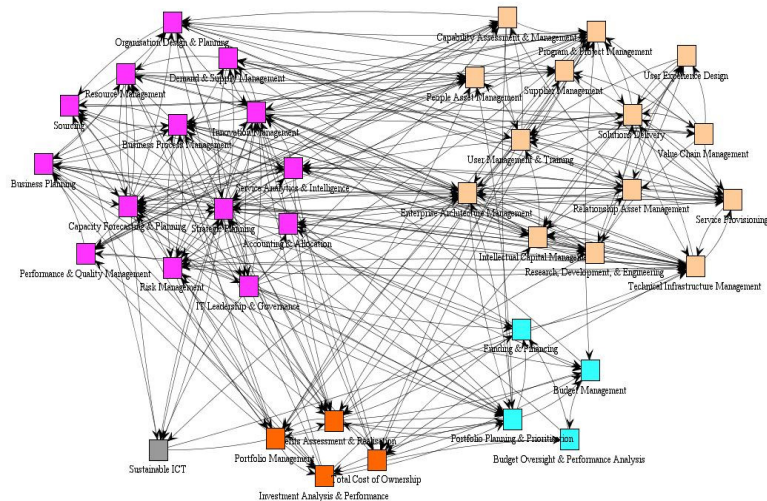


Figure 32: Differentiate groups by nodes colour and location.

## 5.4 Analytical Visualization Features

In this section, I focus on presenting various efforts I have been made to enhance the visualization component of my tool. As I have stated previously, visualization can be effective analysis tool for social network; good visualization may reveal the hidden structure of the networks and amplifies human understanding, thus leading to new insights and new findings. However, visualisation itself cannot serve as an effective and efficient analysis tool for large and complex networks, if it is not equipped with suitable interaction and navigation methods(Herman and Marshall 2000). Graph layout algorithms focus on providing static drawings of relatively large graph and show the entire overview of the graph, which can be effective for revealing patterns and clusters. It is technically possible to display graphs containing millions of nodes or more, but visual clutter, occlusion and other factors can diminish the effectiveness of the visualization as the network complexity increases.

Thus, it is necessary to incorporate various visual analytical techniques to fully support exploration tasks in order to cope with larger networks. Specifically, these visual analytical techniques may include: (1) the use of standard information visualization techniques to augment the network visualization. The simplest augmentation is to map a variable for nodes or edges onto an aesthetic, such as colour, size, shape, pattern, transparency or dashing. Colour, size and shape can be used to depict different characteristic of nodes or edges, which can then be used to explain their structural positions or general network characteristics. The basic art of

labelling nodes is also an example of this use of aesthetics to convey information. (2) The interaction and navigation techniques such as zooming, filtering and distortion can allow viewers to explore only specific sub-regions of interest. Instead of viewing all million nodes, users may get what they need by viewing only the nodes with high out-degree, betweenness centrality and etc. Users may also know facts such as node labels, which they can specify to view only nodes having the given node labels, or their neighbours (Shneiderman and Aris 2006).

Besides these visual analytical techniques, some efforts have also been made to visualize and cluster the network data using multidimensional scaling; allow the users to view the network data on a geographical map. Specifically, in section 5.4.1, I introduce the basic visual analytical techniques such as using colour, shape and label to argument the network visualization. I then cover the interaction and navigation techniques I have implemented in section 5.4.2. In section 5.4.3, I present my work of visualizing and clustering the network data using multidimensional scaling. And in section 5.4.4, I present my work of visualizing the network data on a geographical map.

### **5.4.1 Basic Analytical Visualization**

In social network analysis, in addition to the connections we often have data on attributes of nodes or links of the network; typically in graph visualization we can interpret this kind of information using standard information visualization techniques. For example, by assigning different color, size or shape to each node we can visualize different groups or clusters that each node belongs to as shown in Figure 32. Besides depict different characteristic of nodes or edges, these features can also be used to display the results of different analyses. For example, drawings of networks are often used in combination with an explicit analysis result such as the centrality score as shown in Figure 31. In such a drawing, a graphical feature like the position of a node or the width of an edge could depict the measure under consideration.

Meanwhile, basic graphical features like the color and size of the graph elements are very important for the overall appearance of network drawings. Visualizations based on these features thus offer great potential for creating nice drawings. In my tool,

I have provided interfaces for users to change the visualization properties of the nodes and edges as shown in Figure 26.

## **5.4.2 Interaction and Navigation Techniques**

Unlike other forms of data, visualizing network data not only involves visualizing various attributes for each data point, but it also involves visualizing the links between the data points and the attributes of the links. As a result of trying to show all these pieces of information, visualization of network data are often cluttered, suffer from occlusion and illegible labels, and are difficult to explore. Static drawings can sometimes be helpful, but support for discovery requires an interactive solution to reduce the complexity and enable users to selectively display components of interest. Approaches to an interactive solution involve zooming, filtering, clustering and layout techniques to reduce the number of overlaps and minimize the amount of data to fit in the space available (Namata, et al. 2007).

The Visual Information-Seeking Mantra (Shneiderman 2003) summarizes many visual design guidelines and provides an excellent framework for designing information visualization application: overview first, then zoom and filter, and finally, detail on demand. Keim(Keim 2002) argues in addition to the visualization technique, for an effective data exploration, it is necessary to use some interaction and distortion techniques. Interaction techniques allow the data analyst to directly interact with the visualization and dynamically change the visualization according to the exploration objectives. Distortion techniques help in the data exploration process by providing means for focusing on details which preserve an overview of the data. The basic idea of distortion techniques is to show portions of the data a high level of detail, while others are shown with a lower level of detail. By incorporating these techniques, users can gain a better and more comprehensive understanding of the network data is being explored. Hence I aim to provide plenty of interactions features for users to explore the networks and find actionable insights.

### **5.4.2.1 Zooming**

As the size and complexity of network increases, visualizations of the network are often cluttered. No layout algorithm alone can overcome the problems raised by the

large sizes of the graphs occurring in the visualization applications. Zooming is well-known technique which is widely used in a number of applications. It is quite indispensable when large graph structures are explored. In dealing with large amounts of data, it is important to present the data in a highly compressed form to provide an overview of the data but at the same time allow a variable display of the data on different resolutions (Keim 2002). Zooming is particularly well suited for graphs because the graphics used to display them is usually fairly simple. It is not only a means to display data objects larger but also a means for more details of the data to be represented at higher zoom levels.

In my tool, in the Transform mode it allows users to pan, zoom and scale the graph representation in a more refined way. Users could focus on graphically interesting clusters and start exploring, usually by zooming in to see the connections in a tight group. By scrolling mouse wheel users can zoom in and out for getting detailed views of small areas of the graph. Sometimes zooming into sub-sections of the network may force users to lose the global structure. In order to allow users to gain an overview of the network structure while exploring the detail, I support the satellite view as shown in Figure 28. The satellite view can allow viewers to grasp the overall network structure, in which the full graph is always visible and all mouse actions affect the graph in the master view.

#### **5.4.2.2 Interactive Filtering**

In exploring large data sets, it is important to interactively partition the data set into segments and focus on interesting subsets. This can be done by a direct selection of the desired subset or by a specification of properties of the desired subset. Filtering mechanisms support the produce subgraphs of an original graph. Instead of viewing all the nodes, users may get what they need by viewing only the nodes with high out-degree, betweenness centrality and etc.

Specifically, in my tool I support three types of filtering: nodes, links and K-neighbourhood filter. In the nodes and edges filtering, users can interactively control the visualization to display only a small number of nodes and links, possibly from a very large network. In the implementation, it transforms the input into one which contains only those nodes or links that pass specified predicate. The filtered graph is

a copy of the original graph. In the node filter, only those links whose entire incident node collection passes the predicate are copied into the new graph. In the link filter, all nodes from the original graph are copied into the new graph even if they are not incident to any links in the new graph. The K-neighbourhood filter used to extract the k-neighbourhood around one or more root nodes. The K-neighbourhood is defined as the subgraph induced by the set of nodes that are k or fewer hops away from the root node.

### 5.4.2.3 Interactive Distortion

Interactive distortion techniques support the data exploration process by preserving an overview of the data during drill-down operations. The basic idea is to show portions of the data with a high level of detail while others are shown with a lower level of detail. Distortion techniques are suitable for large but not huge networks. They allow users to place a focus point on a region of interest of the display, and the display redraws so as to magnify the area of interest and de-magnify the rest of the display. Popular versions of these techniques include fisheye and hyperbolic transformations.

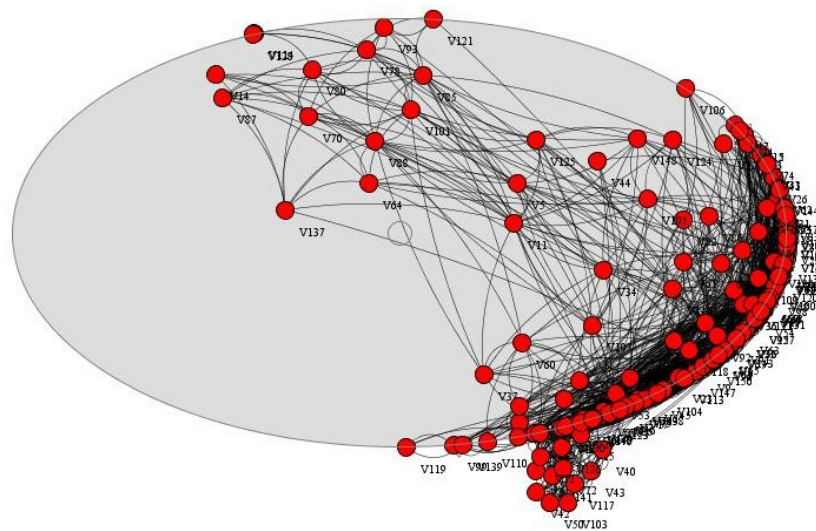


Figure 33: An example of the hyperbolic layout effect.

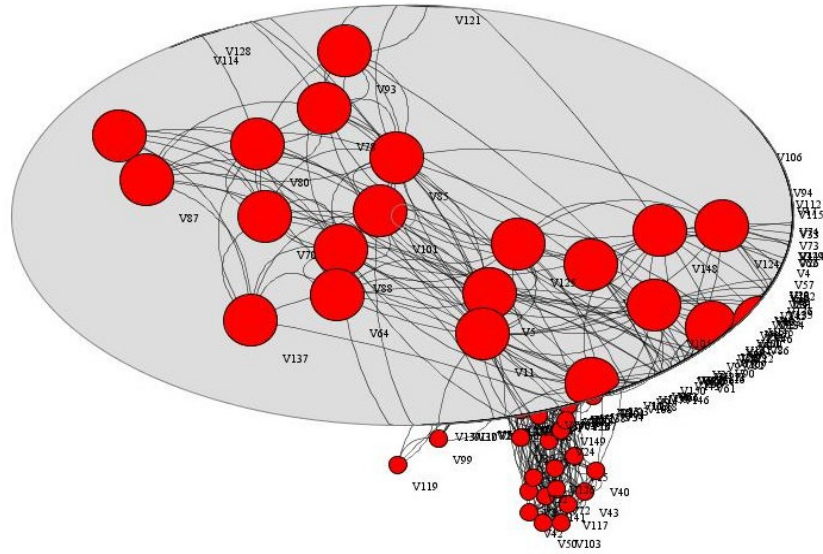


Figure 34: An example of the magnify view effect.

In my tool I support the hyperbolic and magnify transform. In the hyperbolic transform, it creates a fisheye projection of the graph points, with points near the center spread out and points near the edges collapsed onto the circumference of an ellipse. In the magnify transform it creates an enlarging projection of the graph points. Meanwhile, I support both the view and the layout transform. In the view transform it will distort node shapes. Here in Figure 33 I show an example of the hyperbolic layout effect; in Figure 34 I show an example of the magnify view effect.

### 5.4.3 Visualize & Cluster Network Data Using Multidimensional Scaling.

A useful goal for most social network layouts is to represent social distance as physical distance. This representation allows viewers to get a spatial understanding of social relations, as nodes with many relations in common are placed close together on the printed page (Moody, McFarland and Bender-deMoll 2005). Multidimensional scaling (MDS) is a well-known statistical method for mapping pairwise relationships to coordinates that can be used to accomplish this goal. Objects in a data set are represented as points in a geometric space; distance in this space represents proximity or similarity among objects. Objective of MDS is to find coordinates for each point that preserve the given pairwise dissimilarities (distances) as faithfully as possible. It is often used as a tool for understanding relative measurements when absolute measurements are not available (Yang, et al. 2006). A classical MDS

algorithm starts with a matrix of item-item similarities, and then assigns a location to each item in  $N$ -dimensional space, where  $N$  is specified a priori. Formally, it can be described as given the dissimilarity matrix  $D$  of  $n$  points:

$$D_{ij} = d(x_i, x_j) = (x_i - x_j)^T (x_i - x_j) \quad (5.1)$$

The object is to find the coordinates of points  $x_i$  that would give rise to  $D$ .

At the core of MDS is an eigendecomposition on an  $n \times n$  symmetric matrix. Specifically, there are two stages in computing classical MDS. The first is to convert the input matrix  $D$  into a matrix of dot products, or a Gram matrix  $B$ . This is done by multiplying  $D^2$  on both sides with a “centering matrix”  $H$ , which subtracts out the row and column average of each entry and adds back the overall matrix average.

$$B = \frac{-HD^2H}{2} \quad (5.2)$$

The second stage is the bottleneck in MDS. Since  $B$  is symmetric, it can be eigendecomposed into  $USU^T$ , where  $U$  is a matrix of eigenvectors and  $S$  is a diagonal matrix containing the corresponding eigenvalues. MDS derives its lower-dimensional coordinates by taking successive columns from  $U\sqrt{S}$ . A complete eigendecomposition of  $B$  using QR decomposition takes  $O(n^3)$  time, resulting in a  $O(n^3)$  time for MDS. Here I use an example to illustrate the calculation process of classical MDS:

- The input similarity matrix  $D$ : 
$$\begin{bmatrix} 0 & 2 & 5 & 1 \\ 2 & 0 & 4 & 5 \\ 5 & 4 & 0 & 3 \\ 1 & 5 & 3 & 0 \end{bmatrix}$$
- The centred inner product matrix  $B$ : 
$$\begin{bmatrix} 2.5 & 2.375 & -7.5 & 2.625 \\ 2.375 & 6.25 & -1.125 & -7.5 \\ -7.5 & -1.125 & 7.5 & 1.125 \\ 2.625 & -7.5 & 1.125 & 3.75 \end{bmatrix}$$
- Eigendecomposition of  $B$  to get the eigenvectors and eigenvalue matrix: 
$$\begin{bmatrix} 0.6199916017 & 0.5 & 0.4202055682 & -0.4347846528 \\ -0.4299074806 & 0.5 & -0.5212110737 & -0.5417735456 \\ 0.3592287757 & 0.5 & -0.4723094701 & 0.6307760705 \\ -0.5493128969 & 0.5 & 0.5733149757 & 0.3457821278 \end{bmatrix}$$



$$\begin{bmatrix} -5.8181649951 & 0 & 0 & 0 \\ 0 & -0 & 0 & 0 \\ 0 & 0 & 11.5655546358 & 0 \\ 0 & 0 & 0 & 14.2526103594 \end{bmatrix}$$

- The coordinates matrix:

$$\begin{bmatrix} -1.64143 & 1.42904 & NaN & NaN \\ -2.04534 & -1.77254 & NaN & NaN \\ 2.38135 & -1.60624 & NaN & NaN \\ 1.30542 & 1.94974 & NaN & NaN \end{bmatrix}$$

The use of MDS for network visualization has a long history in social network analysis and was first used as such by Laumann and Guttman(1966). MDS based graph layouts are available in common network visualization tools, such as NetDraw, Krackplot and Guess. Substantively, given the measure of similarities or distances between pair of nodes, MDS algorithm calculates the two-dimension coordinates for each node and further plots these nodes. In my implementation, I took advantage of the MDS procedures that are already available in some open source library, and all I need to do is to formalize the node similarity matrix and plot the points with the coordinates given. Here in Table 8 is an example to demonstrate the process of MDS based graph layout, in Figure 35 is the MDS procedure output of a sample Network.

Table 8: Process of visualizing network data using MDS.

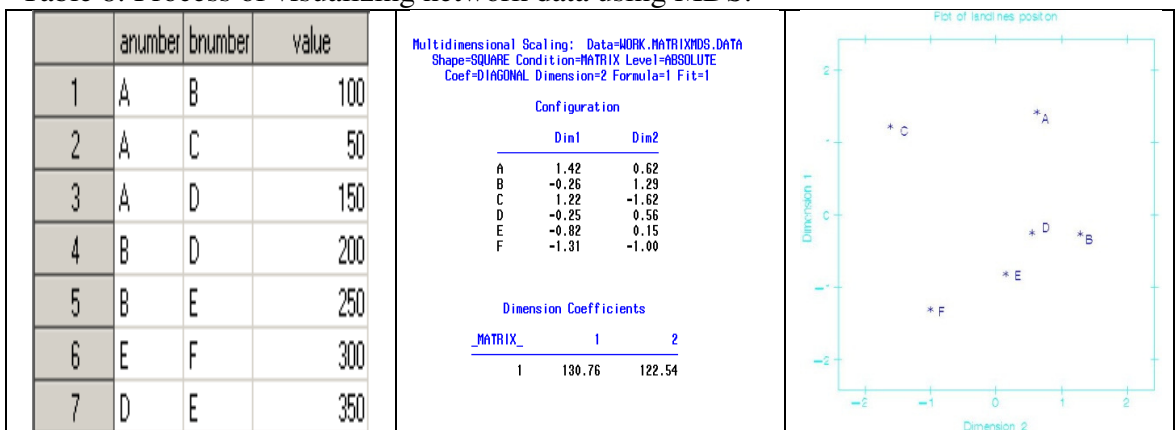




Figure 35: The MDS procedure output of a sample Network.

One of the benefits of using MDS to visualize network data is that it represents social distance as physical distance, which allows the viewers to get a spatial understanding of the social relations by just looking at the image. A social-distance-based representation of network structure is facilitated when edge lengths are equated to relational strengths. An intuitive impression of the network structure can emerge from the proximities in the image.



Figure 36: User interface for setting the clustering options.

Another benefit is that once the visual coordinates are obtained, we can easily cluster the nodes using traditional cluster techniques. Clustering is the unsupervised process of grouping information to achieve a more recognizable presentation of the original data. The computation of a clustering generally requires a distance measure on the data to determine the closeness of data points. In this case, the two-dimension coordinates of each node have already been obtained as the result of the MDS. Hence we can easily calculate the distance of the nodes using common distance

functions such as the Euclidean distance metric. Further, traditional clustering algorithms such as hierarchical clustering can be used to cluster the nodes. In Figure 36 it shows the user interface for setting the clustering options. In Figure 37 is a sample network with three clusters.

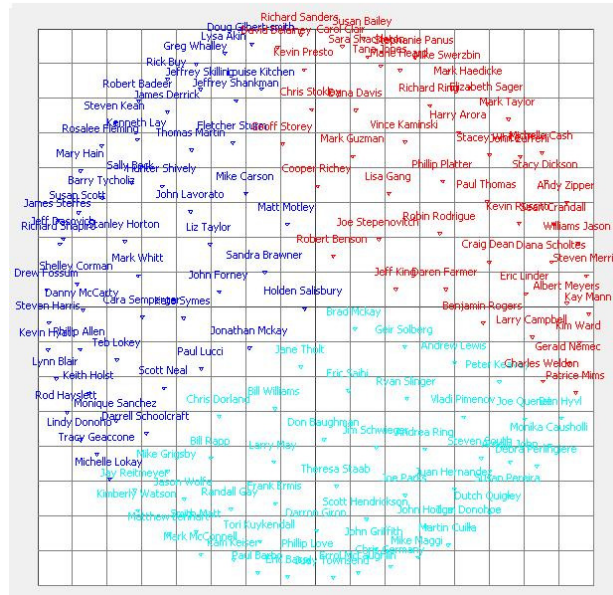


Figure 37: A sample network with three clusters.

This method often produces useful results, although not always for all networks. As stated earlier, a complete eigendecomposition of the centred inner product matrix  $B$  using QR decomposition takes  $O(n^3)$  time, resulting in a  $O(n^3)$  time for MDS. Since the time and space complexity here, application of MDS is limited to instances with at most a few thousands of nodes.

#### 5.4.4 Visualize Network Data on Geographical Map

In some social networks, the spatial information about the nodes may be available. Hence, these networks may have a natural spatial layout such as a geographical trade-flow network or may be abstract as in a personal communications network. In a corporate data warehouse, geographically referenced data such as customer address, street number or zip code may be available. Geographic maps are the most intuitive way to describe and explain the spatial organization of a phenomenon that involves geographically referenced data, that is to say data which has locational references within its structure (Mothe, et al. 2006). Apparently, visualizing social network data

on geographic maps could generate more useful insights than simply visualizing the data.

In many cases, the spatial information we have are in the format of street address or zip code. A geographic coordinate system is a coordinate system that enables every location on Earth to be specified by coordinates. Normally, the combination of latitude and longitude can specify the position of any location on the planet. In order to place markers or position address on a geographic map, sometimes we might need the geographic coordinates of an address. Geocoding is the process of finding associated geographic coordinates often expressed as latitude and longitude from other geographic data, such as street addresses, or zip codes. For example, an address like "1600 Amphitheatre Parkway, Mountain View, CA" can be converted into geographic coordinates like latitude 37.423021 and longitude -122.083739.

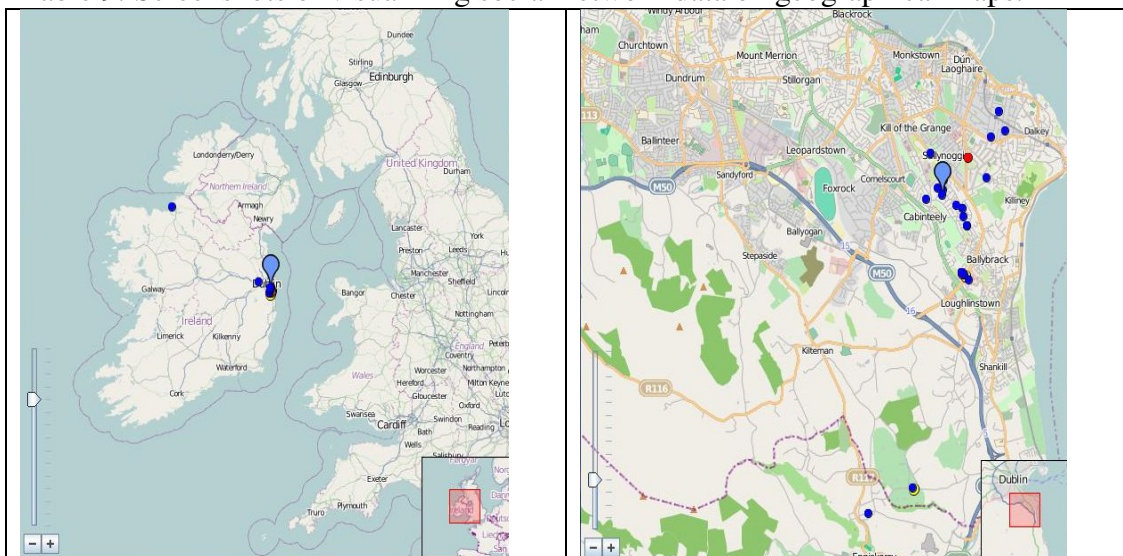
The Google geocoding API provides a direct way to access a geocoder via an HTTP request. Additionally, the service allows you to perform the converse operation (turning coordinates into addresses); this process is known as "reverse geocoding." In my implementation, I use GeoGoogle (<http://geo-google.sourceforge.net/>) a geocoder java API. It is an address standardization API which converting an address into a standardize format. It standardizes addresses by utilizing Google's geocoding service.

After I have got the geographic coordinates for addresses, next thing to do is to put these locations on a map. I aim to find open source libraries that can embed mapping abilities into Swing-based java applications. The JXMapView is an open source Swing component created by the developers at SwingLabs. At its core, the JXMapView is a special JPanel that knows how to load tiles from an image server. All of the details of how to convert coordinate to pixels, cache tiles, and stitch them together on screen are nicely hidden inside JXMapView's API. All users need to do add it to the Swing application the way like any other JPanel. A kit version of JXMapView includes zoom buttons, a zoom slider, and a mini-view. The JXMapKit is just a wrapper for the JXMapView that includes the most commonly used features. Users can drag the map around and also zoom in and out to different levels. The JXMapView comes preconfigured with connections to Open Street

Map and the Blue Marble, but users can use a custom tile provider to connect to an alternate map server such as Google maps.

To make a geographical map really useful, points that represent locations must be drawn on it. In the mapping world, coordinates that represent physical locations are called waypoints. In network visualization, nodes are usually represented as points, and links among nodes are represented by lines. Similarly, when it comes to visualizing network data on a geographical map, we can use waypoints to represent the nodes and draw lines among nodes to represent the links. By default, waypoints are drawn using a blue teardrop shape, but users can custom the looks of waypoints by using the waypoint painter. Painters are classes that implement the painters interface and can be set on a SwingX component to customize that component's drawing. Easily we can change the size, colour and shapes of waypoints. Besides custom the looks of waypoints, we can also custom the overlays to draw some lines or polygons between waypoints. Herein Table 9 are some examples with different zoom levels:

Table 9: Screenshots of visualizing social network data on geographical maps.



## **Chapter 6**

### **Conclusion and Future Work**

In this chapter, the conclusion of the thesis is given. I present the summary of my work, its contributions, limitations and future work direction. Specifically, in section 6.1, I give a summary of my work and explain how this research contributes to both theory and practice. In section 6.2, I discuss the limitations of my work. In section 6.3, the future work direction for this research is given.

#### **6.1 Summary & Contributions**

In this section, I first present the summary of my work then discuss its contributions.

##### **6.1.2 Summary of Work**

There is a growing interest in exploring the role of social networks for understanding how individuals spread influence. In customer intelligence, social network analysis is fast emerging as an important discipline for predicting and influence consumer behaviour. Within online social networking, social media plays a prominent role in promotion, marketing and public relations. Hence, it is important to gain some deep insights of how information diffuses through social networks.

In this context, in chapter 2, I start with an overview of the social network analysis research with a focus on the network centrality measures and cohesive subgroup related concepts and algorithms. I then introduce various state of the art research issues and provide theoretical background for my work on modelling diffusion process and empirical diffusion studies. After that, I provide an overview of some of the major social network analysis tools, summarizing their advantages and disadvantages. I also introduce various graph layout algorithms that used to position the nodes and edges in the network visualization.

In chapter 3, I present my work of modelling the diffusion processes in social networks. I extend an epidemic model that accurately models virus propagation to

describe the diffusion process in social networks. With proper parameters setting on a given network topology, myproposed method can numerically calculate each node's probability to get infected when a set of nodes has been initially activated. By comparing its predicting performance with diffusion simulations, I validate the accuracy of my proposed method in capturing diffusion process. Further, using the number of nodes reached by set of initial active nodes as an influence measure, experiment results show that my proposed method provides ways of extracting nontrivial nodes as influential nodes.

In chapter 4, I present my work on empirically measuring the diffusion processes. I obtain data from a telecommunication operator and construct customer call networks using the call records data. I then analyze the call networks and assess social influence in the bundle adoption process. Specifically, the analysis results show that the call networks are relatively stable during the examination period. Using correlation analysis I confirm the existence of social influence in the call networks with regard to bundle adoption event. However, the characteristic of the influence is not same as it has been observed in other scenarios.

In chapter 5, I present my work on the prototypical tool for social network data manipulation, analysis and visualization. The tool has rich support for dynamically manipulating and interactively exploring social network data. It also supports network visualization techniques such as satellite view, "lens effect" and visualizing social network data on a geographical map.

## **6.1.2 Contributions**

As I have mentioned in the introduction chapter, this thesis offers three principal contributions to both theory and practice.

The wide variety of rules theoretical diffusion models posed on individuals' behaviour, even if plausible, are often lacking empirical support (Cointet and Roth 2007). It is necessary to propose new models that can capture the diffusion phenomenon more accurately. Here I attempt to extend the classic diffusion models to integrate some of the recent empirical findings on the diffusion phenomenon. I propose a new method to approximate the diffusion processes based on a Non-linear

Dynamic System that describing disease spread in epidemiology. With proper parameters setting on a given network topology, the proposed method can numerically calculate each node's probability to get infected when a set of nodes has been initially activated. It provides an alternative way of estimating the number of nodes reached by the initial target set in the diffusion process. Using this quantity as an influence measure, experiment results show that the proposed method can provide a way of identifying nontrivial nodes as influencer.

Collecting social network data has traditionally been hard work. The emergence of social media has transformed the communication patterns among people. With the availability of such large-scale real network data, we can now measure the diffusion phenomenon at a more quantitative level. Indeed, it is important to obtain real world network data and conduct empirical analysis on such processes. Here I conduct an empirical study that assesses social influence in the telecommunication call networks with regard to the bundle adoption event. The analysis results show that the call networks are relatively stable during the examination period. Using correlation analysis the study confirms the existence of social influence in the call networks with regard to bundle adoption event. However, the characteristic of the influence is not same as it has been observed in other scenarios.

Many of the network analysis tools are designed for expert practitioner, have complex data handling, and inflexible graphing and visualization features that inhibit wider adoption. Meanwhile, they have extensive and complicated analysis and visualization features. Hence, I believe it is necessary to come up with a simple tool that can allow users with little knowledge on social network analysis can perform analysis tasks. Here I develop a prototypical tool for social network data manipulation, analysis and visualization. The tool offers extensive features for users to interactively manipulate and explore social networks. It also provides distinctive visualization features such as allowing users to view the network data on a geographical map.

## **6.2 Limitations**

In this section, I discuss the limitations of the research. With regard to the three principal contributions of the thesis, the limitations are as follows:



Even though the method I proposed to capture the diffusion process in social networks has been proven can give a good approximation, it does not outperform traditional methods as I expected. On the other hand, I aim to propose new models that can capture the diffusion phenomenon more accurately. It is unclear to what extent the proposed method has rendered the diffusion phenomenon. There is still much remains unknown regarding to the diffusion phenomenon.

With regard to the empirical studies I conduct on the diffusion process using the broadband bundle adoption data to assess social influence in the call networks, the results somehow confirm the existence of social influence in the call networks, as the friends of the early adopters are more likely to also adopt the bundle. However, when investigating how does one's probability of adopting depending on the number of friends who have already adopted, the results are not as expected that the probability keeps increases as more friends adopted, as it has been observed in many other scenarios.

With regard to the tool I developed to visualize, analyze and manipulate social network data, it is still very prototypical - many functionalities and features are remain to be included.

### **6.3 Future Work Direction**

Based on the limitations of the research, the future work directions include:

With regard to my work on modelling diffusion process, my future work directions include obtaining actual information diffusion data and therefore I could develop ways to infer or estimate relevant model parameters with the historical diffusion data. With the support of the empirical findings I could make more general assumptions on how individuals respond to friends' influence, which leads to a closer integration of the theoretical models to the empirical results. Meanwhile, as my method has not outperformed traditional diffusion simulation as I expected, in the future I need to improve the calculation process so that it can be fast.

With regard to the empirical studies on diffusion process, in the future work I would like to obtain more detail data that allow us to better capture the underlying social interactions within the phone users. Meanwhile, I would like to investigate other

business events such as customer churn to assess social influence and measure the diffusion process, to see if the same kind of characteristics is exhibited in other scenarios.

With regard to the network analysis and visualization tool, my future work include:

(1) Test the error handling and scalability of the tool (2) Support other network data format. (3) Enhance the interactively exploration and navigation features.

## References

- Aaron Quigley. Large Scale Force Directed Layout [Online]. Available from: <http://www.cs.usyd.edu.au/~aquigley/3dfade/>.
- Adar, E. 2006. Guess: a language and interface for graph exploration. *IN: Proceedings of the SIGCHI conference on Human Factors in computing systems*. pp213-222, ACM.
- Anagnostopoulos, A., Kumar, R. and Mahdian, M. 2008. Influence and correlation in social networks. *IN: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp7-15, ACM.
- Anderson, R.M. and May, R.M. 1992. *Infectious diseases of humans: dynamics and control*. Oxford University Press, USA.
- Au, S. C., Leckie, C., Parhar, A. and Wong, G. 2004. Efficient visualization of large routing topologies. *International Journal of Network Management*. 14 (2), pp105-118.
- Backstrom, L., Huttenlocher, D., Kleinberg, J. and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. *IN: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.
- Becker, R. A., Eick, S. G. and Wilks, A. R. 1995. Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics*. 1 (1), pp16-28.
- Bollobás, B. 1984. The evolution of random graphs. *Transactions of the American Mathematical Society*. 286 (1), pp257-274.
- Brandes, U. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*. 25 (2), pp163-177.
- Brandes, U., Kenis, P. and Raab, J. 2006. Explanation through network visualization. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2 (1), pp16-23.
- Brandes, U., Kenis, P., Raab, J. 1999. Explorations into the visualization of policy networks. *Journal of Theoretical Politics*. 11 pp75-106.
- Brandes, U., Raab, J. and Wagner, D. 2001. Exploratory network visualization: Simultaneous display of actor status and connections. *Journal of Social Structure*. 2 (4), pp1.
- Burt, R.S. 2005. *Brokerage and closure: An introduction to social capital*. Oxford University Press, USA.

- Carnes, T., Nagarajan, C., Wild, S.M. and van Zuylen, A. 2007. Maximizing influence in a competitive social network: a follower's perspective. *IN: Proceedings of the ninth international conference on Electronic commerce*. ACM.
- Cha, M., Mislove, A. and Gummadi, K.P. 2009. A measurement-driven analysis of information propagation in the flickr social network. *IN: Proceedings of the 18th international conference on World wide web*.pp721-730. ACM.
- Chakrabarti, D., Wang, Y., Wang, C. 2008. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*. 10 (4), pp1.
- Chan, D., Chua, K.S., Leckie, C. and Parhar, A. 2003. Visualisation of powerlaw network topologies. *IN: Proc. of the 11th IEEE Intl. Conf. on Networks*.pp69-74.Citeseer.
- Christakis, N. A. and Fowler, J. H. 2007. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*. 357 (4), pp370.
- Cointet, J. P. and Roth, C. 2007. How realistic should knowledge diffusion models be? *Journal of Artificial Societies and Social Simulation*. 10 (3), pp5.
- Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. *IN: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.
- Cross, R., Thomas, R.J., 2009. Driving results through social networks: how top organizations leverage networks for performance and growth. Jossey-Bass.
- Dasgupta, K., Singh, R., Viswanathan, B. 2008. Social ties and their relevance to churn in mobile telecom networks. *IN: Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM.
- de Kerchove, C., Krings, G., Lambiotte, R. 2009. Role of second trials in cascades of information over networks. *Physical Review E*. 79 (1), pp16114.
- Derényi, I., Palla, G. and Vicsek, T. 2005. Clique percolation in random networks. *Physical Review Letters*. 94 (16), pp160202.
- Ding, C., He, X., Husbands, P., Zha, H. and Simon, H.D. 2002. PageRank, HITS and a unified framework for link analysis. *IN: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.
- Domingos, P. and Richardson, M. 2001. Mining the network value of customers. *IN: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

- Doyle, S. 2008. Software review Social network analysis in the Telco sector—Marketing applications. *Journal of Database Marketing & Customer Strategy Management*. 15pp130-134.
- Du, N., Wu, B., Pei, X., Wang, B. and Xu, L. 2007. Community detection in large-scale social networks. *IN: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*.pp16-26. ACM.
- Dunn, R., Dudbridge, F. and Sanderson, C. M. 2005. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC bioinformatics*. 6pp39.
- Eades, P. 1984. A heuristic for graph drawing. *Congressus numerantium*. 42 (149160), pp194-202.
- Estevez, P., Vera, P. and Saito, K. 2007. Selecting the most influential nodes in social networks. *IN: Neural Networks, 2007. IJCNN 2007*. pp2397-2402
- Farahat, A., LoFaro, T., Miller, J. C. 2006. Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*. 27 (4), pp1181-1201.
- Freeman, L.C. 2004. *The development of social network analysis*. Empirical Press Vancouver, British Columbia.
- Freeman, L. C. 2000. Visualizing social networks. *Journal of social structure*. 1 (1), pp4.
- Freeman, L. C. 1979. Centrality in social networks conceptual clarification. *Social networks*. 1 (3), pp215-239.
- Fruchterman, T. M. J. and Reingold, E. M. 1991. Graph drawing by force-directed placement. *Software- Practice and Experience*. 21 (11), pp1129-1164.
- Gabbott, M. and Hogg, G. 1994. Consumer behaviour and services: a review. *Journal of Marketing Management*. 10 (4), pp311-324.
- Gansner, E. and Koren, Y. 2006. Improved circular layouts. *IN: Graph Drawing*. Springer.
- Girvan, M. and Newman, M. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. 99 (12), pp7821.
- Goldenberg, J., Libai, B. and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*. 12 (3), pp211-223.
- Granovetter, M. 1978. Threshold models of collective behavior. *The American Journal of Sociology*. 83 (6), pp1420.

- Granovetter, M. S. 1973. The strength of weak ties. *The American Journal of Sociology*. 78 (6), pp1360.
- Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. 2004. Information diffusion through blogspace. *IN: Proceedings of the 13th international conference on World Wide Web*. ACM New York, NY, USA.
- Hanneman, R. A. and Riddle, M. 2005. Introduction to social network methods. *Riverside, CA: University of California, Riverside*.
- Hartline, J., Mirrokni, V. and Sundararajan, M. 2008. Optimal marketing strategies over social networks. *IN: Proceeding of the 17th international conference on World Wide Web*. ACM.
- Herman, I. and Marshall, M. S. 2000. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*. 6(1): pp24-43
- Hidalgo, C. and Rodriguez-Sickert, C. 2008. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017-3024.
- Huisman, M. and Van Duijn, M. A. J. 2005. Software for social network analysis. *Models and methods in social network analysis*. pp270-316.
- Immorlica, N., Kleinberg, J., Mahdian, M. and Wexler, T. 2007. The role of compatibility in the diffusion of technologies through social networks. *IN: Proceedings of the 8th ACM conference on Electronic commerce*. ACM.
- Kamada, T. and Kawai, S. 1989. An algorithm for drawing general undirected graphs. *Information processing letters*. 31 (1), pp7-15.
- Katona, Z., Zubcsek, P. and Sarvary, M. 2009. Network effects and personal influences: diffusion of an online social network. .Working Paper.
- Katz, E., Lazarsfeld, P.F. 1955. Personal influence: the part played by people in the flow of mass communications. Glencoe, IL: Free Press.
- Keim, D. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*. 8 (1), pp1-8.
- Kempe, D., Kleinberg, J. and Tardos, É. 2005. Influential nodes in a diffusion model for social networks. *Automata, Languages and Programming*. pp1127-1138.
- Kempe, D., Kleinberg, J. and Tardos, É. 2003. Maximizing the spread of influence through a social network. *IN: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA.
- Kleinberg, J. 2008. The convergence of social and technological networks. *Communications of the ACM*. 51 (11), pp66-72.

- Kleinberg, J. 2007. Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic game theory*.
- Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*. 46 (5), pp604-632.
- Kossinets, G., Kleinberg, J. and Watts, D. 2008. The structure of information pathways in a social communication network. *IN: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.
- Laumann, E. O. and Guttman, L. 1966. The relative associational contiguity of occupations in an urban setting. *American Sociological Review*. 31 (2), pp169-178.
- Lazarfeld, P.F., Berelson, B. and Gaudet, H. 1944. *The people's choice: how the voter makes up his mind in a presidential campaign*. Columbia Univ Pr.
- Leskovec, J., Adamic, L. A. and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*. 1 (1), pp5.
- Leskovec, J., McGlohon, M., Faloutsos, C. 2007. Cascading behavior in large blog graphs: Patterns and a model. *Society of Applied and Industrial Mathematics: Data Mining*.
- Liben-Nowell, D. and Kleinberg, J. 2008. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*. 105 (12), pp4633.
- Mahajan, V., Muller, E. and Bass, F. M. 1990. New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing*. 54 (1), pp1-26.
- McPherson, M., Smith-Lovin, L. and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*. 27 (1), pp415-444.
- Meyer, B. 1998. Self-organizing graphs—a neural network perspective of graph layout. *IN: Graph Drawing*. V.1547/1998, pp246-262. Springer.
- Milgram, S. 1967. The small world problem. *Psychology today*. 2 (1), pp60-67.
- Moody, J., McFarland, D. and Bender-deMoll, S. 2005. Dynamic network visualization1. *The American Journal of Sociology*. 110 (4), pp1206-1241.
- Morris, S. 2000. Contagion. *Review of Economic Studies*. 67 (1), pp57-78.
- Mothe, J., Chrisment, C., Dkaki, T. 2006. Combining mining and visualization tools to discover the geographic structure of a domain. *Computers, Environment and Urban Systems*. 30 (4), pp460-484.

- Namata, G.M., Staats, B., Getoor, L. and Shneiderman, B. 2007. A dual-view approach to interactive network visualization. *IN: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM.
- Navarro, J. F., Frenk, C. S. and White, S. D. M. 1997. Hierarchical Clustering. *The Astrophysical Journal*. 490pp493-508.
- Newman, M. 2007. Mathematics of networks. *In The New Palgrave Encyclopedia of Economics, 2nd edition*. Palgrave Macmillan, Basingstoke.
- Newman, M. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*. 103 (23), pp8577.
- Newman, M. 2004. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*. 38 (2), pp321-330.
- O'Madadhain, J., Fisher, D., Smyth, P. 2005. Analysis and visualization of network data using JUNG. *Journal of Statistical Software*. 10pp1-35.
- Onnela, J. P., Saramäki, J., Hyvönen, J. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*. 104 (18), pp7332.
- Page, L., Brin, S., Motwani, R. and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. .
- Palla, G., Derényi, I., Farkas, I. and Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*. 435 (7043), pp814-818.
- Perer, A. and Shneiderman, B. 2008a. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. .
- Perer, A. and Shneiderman, B. 2008b. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. *IN: Proc. 13th Int'l Conf. on Intelligent User Interfaces*. Citeseer.
- Perer, A. and Shneiderman, B. 2006. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*. 12 (5), pp693-700.
- Porter, M. A., Onnela, J. P. and Mucha, P. J. 2009. Communities in networks. *Notices of the American Mathematical Society*. 56 (9), pp1082-1097.
- Purchase, H. C. 2002. Metrics for graph drawing aesthetics. *Journal of Visual Languages & Computing*. 13 (5), pp501-516.
- Purchase, H. C., Carrington, D. and Alder, J. A. 2002. Empirical evaluation of aesthetics-based graph layout. *Empirical Software Engineering*. 7 (3), pp233-255.



- Reingold, E. M. and Tilford, J. S. 1981. Tidier drawings of trees. *IEEE Transactions on Software Engineering*. 7 (2), pp223-228.
- Rogers, E.M. 1995. *Diffusion of innovations*. Free Pr.
- Saito, K., Nakano, R. and Kimura, M. 2008. Prediction of information diffusion probabilities for independent cascade model, *KES* (3), pp. 67-75
- Seidman, S. B. 1983. Internal cohesion of ls sets in graphs\* 1. *Social Networks*. 5 (2), pp97-107.
- Shneiderman, B. 2003. The eyes have it: A task by data type taxonomy for information visualizations. *The craft of information visualization: readings and reflections*. pp364-371.
- Shneiderman, B. and Aris, A. 2006. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics*. 12 (5), pp733-740.
- Smith, M.A., Shneiderman, B., Milic-Frayling, N. 2009. Analyzing (social media) networks with NodeXL. *IN: Proceedings of the fourth international conference on Communities and technologies*. ACM.
- Song, X., Chi, Y., Hino, K. and Tseng, B.L. 2007. Information flow modeling based on diffusion rate for prediction and ranking. *IN: Proceedings of the 16th international conference on World Wide Web*. ACM.
- Tamassia, R. (ed.) 2006. *Handbook of graph drawing and visualization*. Chapman & Hall/CRC.
- Team, N. 2006. *Network Workbench Tool*. Indiana University, Northeastern University, and University of Michigan.
- Tufte, E.R. and Howard, G. 1983. *The visual display of quantitative information*. Graphics Press Cheshire, CT.
- Tunkelang, D. 1999. A Numerical Optimization Approach to General Graph Drawing. .
- Van den Bulte, C. and Stremersch, S. 2004. Social contagion and income heterogeneity in new product diffusion: A meta-analytic test. *Marketing Science*. 23 (4), pp530-544.
- Viégas, F.B. and Donath, J. 2004. Social network visualization: Can we go beyond the graph. *IN: Workshop on Social Networks, CSCW*. Citeseer.
- Wagner, D. 2003. Analysis and Visualization of Social Networks. *IN: Experimental and Efficient Algorithms: Second International Workshop, WEA 2003, Ascona, Switzerland, May 26-28, 2003. Proceedings*. Springer.

- Wasserman, S. and Faust, K. 2007. *Social network analysis: Methods and applications*. Cambridge Univ Pr.
- Watts, D. J. and Dodds, P. S. 2007. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441-458.
- Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature*. 393 (6684), pp440-442.
- Wu, F. and Huberman, B. 2004. Finding communities in linear time: a physics approach. *The European Physical Journal B-Condensed Matter and Complex Systems*. 38 (2), pp331-338.
- Yang, T., Liu, J., McMillan, L. and Wang, W. 2006. A fast approximation to multidimensional scaling. *IN: IEEE workshop on Computation Intensive Methods for Computer Vision*. Citeseer.