Social Network Analysis Based on a Hierarchy of Communities

Xiao Cui

College of Engineering and Science
Victoria University

Submitted in fulfillment

of the requirements of the degree of

Doctor of Philosophy

March, 2016

Abstract

With the rapid growth of users in Social Networking Services (SNSs), data is generated in thousands of terabytes every day. This data contains lots of hidden information and patterns. The analysis of such data is not a trivial task. A great deal of effort has been put into it. Analysing users' behaviour in social networks can help researchers to better understand what happens in the real world and create huge commercial value for social networks themselves.

Research on Social Network Analysis (SNA) includes a wide range of topics such as network modelling, centrality analysis, community detection, etc. Many of these topics have been well-studied and some of them have found practical use in real-world problems. However, most of these studies rely on an implicit assumption, that social networks are flat. There are few studies analysing social networks at different levels of abstraction.

In this research work, a model called Social Network Analysis based on a Hierarchy Of Communities (SNAHOC) was designed to study social networks at different levels of abstraction. The ultimate goal was to find the hidden information and patterns in social networks that are not obtainable through classical approaches. Case studies, based on SinaData, were conducted to examine the capability of the SNAHOC model. The first case study examined the influence of geographic diversity on network topology. The degree of a community is positively related to the variety of locations the community has. The second case study explored the factors that can have an impact on the influence of a community. Prominent users do have a substantial impact on the influence of a community but it is not the only determinant. Both studies benefited by the multilevel analysis of SNAHOC.

This thesis also discusses the methods for data preprocessing, with respect to SinaData, which was retrieved from Sina Weibo, by the crawlers written for this thesis.

Student Declaration

I, Xiao Cui, declare that the PhD thesis entitled "Social Network Analysis Based on A Hierarchical Model of Community Structure" is no more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

Signature: [redacted]

Date: 10/03/2016

Acknowledgments

I would like to express my special appreciation and thanks to both of my supervisors, Associate Professor Hao Shi and Professor Xun Yi, for their help, advice, and mentoring during my study. Without their support and guidance, none of the work presented in this thesis would have been possible.

I would like to express my deep gratitude to my wife Yao Qu and all of my family members, for their love and support during the ups and downs of my study. I would also take this opportunity to thank those who have directly and indirectly helped me.

Publications

Publications during the study:

Xiao Cui, Hao Shi, and Xun Yi. Application of Association Rule Mining Theory in Sina Weibo. *Journal of Computer and Communications*, 14(2):19-26, 2014.

Xiao Cui, and Hao Shi. An Overview of Pathfinding in Navigation Mesh. *International Journal of Computer Science and Network Security*, 12(12):48, 2012.

Xiao Cui, and Hao Shi. Direction Oriented Pathfinding in Video Games. *International Journal of Artificial Intelligence & Applications*, 11(2), 2011.

Xiao Cui and Hao Shi. A*-based Pathfinding in Modern Computer Games. *International Journal of Computer Science and Network Security*, 11(1):125-130, 2011.

Contents

List of Figures

List of Tables

Glossary

API A set of functions and procedures that allow the creation of applications which access the features or data of an operating system, application, or other service.

Community A cluster of users who are densely connected internally.

Community ID A unique identifier for a community.

Community level A graph where vertices represent communities and edges represent connections between the communities.

D1A A directed graph derived from Sample 1A.

D2A A directed graph derived from Sample 2A.

Degree of a community The number of connections of a community with other communities.

Degree of a vertex The number of edges incident to a vertex.

Evenness A measure of the equality or distribution of locations in a cluster.

Followee Someone who is being tracked on a social network.

Follower Someone who is tracking a particular person, group, organization, etc. on a social network.

Friend Someone who follows you back when you follow him/her

Geographic diversity The variety of locations in a cluster. Measured by Shannon-Wiener index, that incorporates both richness and evenness.

Indegree of a community The number of connections directed at a community.

Indegree of a vertex The number of edges directed at a vertex.

Inter-connection A connection between two entities who are in different clusters.

Intra-connection A connection between two entities who are in the same cluster.

Isolated community A community that has no connections with any other communities.

Isolated user Someone who is isolated from any other users.

LCC1A The largest connected component of U1A.

LCC1A-1 A graph which is an abstraction of LCC1A at community level.

LCC1A-1-1 An equivalent of LCC1A-1. In order to show the distribution of locations in communities, vertices are replaced by pie-charts.

LCC1A-1-2 An equivalent of LCC1A-1 where vertices are labelled by their richness.

LCC1A-1-3 An equivalent of LCC1A-1 where vertices are labelled by their geographic diversity.

LCC1A-2 A graph which is an abstraction of LCC1A at society level.

LCC1A-2-2 Generated by LCC1A-2 and LCC1-A-1-2, where communities are bounded by societies and the size of vertices varies depending on their richness.

LCC1A-2-3 Generated by LCC1A-2 and LCC1-A-1-3, where communities are bounded by societies and the size of vertices varies depending on their geographic diversity.

Monthly active user A key performance indicator for social networks. Calculated by taking the number of unique users (such as someone who logged in at least once) within the previous 30 days.

OAuth An open standard for authorisation. OAuth provides client applications a secure delegated access to server resources on behalf of a resource owner.

Outdegree of a community The number of connections directed out of a community.

Outdegree of a community The number of edges directed out of a vertex.

Pioneer Someone who is active in Sina Weibo.

RD2A A directed graph generated by removing isolated vertices in D2A.

RD2A-1 A graph which is an abstraction of RD2A at community level.

RD2A-2 A graph which is an abstraction of RD2A at society level.

REST API A set of functions which developers can perform requests and receive responses via HTTP protocol such as GET and POST.

Richness The number of different locations in a cluster.

SWI Shannon-Wiener Index. A measure of the geographic diversity of a community.

Society ID A unique identifier for a society.

Social connection A connection between two entities, either in an asymmetric way (directed) or a symmetric way (undirected).

Social diversity A measure of the sphere of the influence of an entity. Calculated by the number of inter-connections of a vertex.

Social network A dedicated website or other application which enables users to communicate with each other by posting information, comments, messages, images, etc.

Society A community of communities who are densely connected internally.

Society level A graph where vertices represent societies and edges represent connections between the societies.

U1A An undirected graph transformed from D1A.

Unverified Someone whose real identity has not been verified yet.

VIP Someone whose real identity is a celebrity, corporation, government department, etc.

We-media	Grassroots Internet journalists.
Weibo	A message sent in Sina Weibo.

Chapter 1

Introduction

The use of Social Networking Services (SNSs) has exploded in the past decades, that has created big changes in everyday communications. With the widespread availability of wireless Internet access and the rapid development of mobile devices, people are allowed to access to SNSs at any time, any place. Analysing users' behaviour in SNSs can help us to understand what happens in the real world (Alef, 2010).

A graph is often used to model a social network, where vertices represent individual users and edges represent the social connections between the users. Social networks show strong community structure (Girvan and Newman, 2002), in which vertices are relatively densely connected within groups but sparsely connected between groups. The techniques for community detection have been well-studied (Clauset *et al.*, 2004; Blondel *et al.*, 2008; Pons and Latapy, 2005; Raghavan *et al.*, 2007; Newman, 2004; Hui *et al.*, 2007; Mucha *et al.*, 2010; Sun *et al.*, 2009; Cafieri *et al.*, 2014). However, few researchers explore the use of community structures for Social Network Analysis (SNA). Many papers (Benevenuto *et al.*, 2009; Kwak *et al.*, 2010; Gyarmati and Trinh, 2010; Yan *et al.*, 2013; Wu *et al.*, 2011) discuss the characteristics of users in social networks but few papers examine the characteristics of communities in social networks.

1.1 Research Objectives

The aim of this research work was to develop a model to examine social networks at different level of abstraction using a hierarchy of communities, and formulate Social Network Analysis based on a Hierarchy Of Communities (SNAHOC) taking into consideration the properties of large-scale networks.

This research work also developed a set of tools for data collection and data pre-processing using Sina Weibo data, in order to conduct visual analysis of massive volumes of data.

In addition, in order to explore the hidden patterns in communities, a set of aggregation functions, for the measurement of geographic diversity of communities and the influence of communities were, developed. This research work used SNAHOC to conduct case studies to identify the influence of geographic diversity on network topology and the factors that can have an impact on the influence of a community.

This research work represents the application of a multilevel model to SNA, in which the networks are explored by SNAHOC at different hierarchical abstraction levels.

1.2 Research Methods

This research work involved seven stages: developing the SNAHOC model, selecting a data source, collecting data, pre-processing data, designing case studies, sampling data and demonstrating the use of SNAHOC with respect to the case studies.

Stage one involved the development of the concepts used in SNAHOC. The concept of levels of abstraction was introduced so that social networks can be viewed at different scale level. Stage two involved the selection of a data source. Sina Weibo was selected because it contains rich information of users, has not been well studied, is easy to access via official API and most importantly, users' real identities have already been verified. Stage three involved collecting data from the selected data source. A crawler program that specifically target Sina Weibo was written in Java.

In stage four, the data set, SinaData, was cleaned up. In stage five, two case studies were designed to demonstrate the capability of SNAHOC. Stage six involved drawing samples taking into consideration the nature of the case studies. In stage seven, the use of SNAHOC was demonstrated. Networks were organised into a hierarchy of communities and the information was aggregated every time passing the networks to a higher level of abstraction.

1.3 Scope of the Research

This research work is presented in eight chapters. Chapter 1 provides an introduction to the research project highlighting the research background, objectives and research scope. Chapter 2 gives the context in which SNAHOC works by delineating the main lines of research on SNA. Chapter 3 reviews the mathematical theories behind SNA and also presents the techniques used in the case studies.

Chapter 4 outlines the steps in designing SNAHOC. The concepts of SNAHOC are given in plain English and explained by graph theory as well.

Chapter 5 discusses data collection, data pre-processing and the methods for data sampling. The reason for using Sina Weibo as the data source is given. The methods used for data collection are discussed. The collected data is named SinaData and the data structure of SinaData is defined. Finally, the need for sampling is discussed and related tools for sampling are given.

Chapter 6 conducts a case study to investigate the influence of the geographic diversity of a community on the social diversity of the community. The relationship between the social diversity of a community and the factors determining the geographic diversity of the community are examined based on SNAHOC. At the end of the chapter, the experimental results are compared with existing social theory.

Chapter 7 presents a measure of the influence of a community based on SNAHOC. The characteristics of the communities that are considered influential are also discussed. The relationship between individual users and communities in terms of influence are examined.

Chapter 8 summaries the research work and makes suggestions for further work. It concludes that viewing social networks at different levels of abstraction can successfully extract the information and patterns that are not obviously observable in a ‘flat world’.

Chapter 2

Literature Review

This chapter delineates the main lines of research on SNA in the following order: first, the definitions about social networks are presented, then the most basic characteristics of social networks are described, following by the major challenges that SNA faces, at the end, the main lines of research on SNA are discussed. This chapter gives the context for where Social Network Analysis based on a Hierarchy Of Communities (SNAHOC) works.

2.1 Social Network

To ordinary people, websites such as Facebook, Twitter, and LinkedIn have become synonymous with social networks. The people who use websites like Facebook and Twitter are called ‘users’ and the websites themselves are categorized as ‘Social Networking Sites’ or ‘Social Networking Services’ (both abbreviated to SNSs). Sociologists use different terms, e.g. they prefer to use “social actors” instead of users, but considering the inextricable connection between social networks and the Internet, most researchers in computer science prefer the terms that have been widely used on the Internet rather than the jargon of sociology.

Social networks can exist in different forms. Facebook is a typical social networking service where people are allowed to build and maintain their social connections with other people of similar interests and background. Communication on Facebook is based on a two-way relationship, e.g. user A alone can not build a connection with user B unless user B wants to build a connection with user A as well. Unlike Facebook, relationships on Twitter are based on an asymmetric model, e.g. user A can 'follow' user B, whether user B follows user A or not. News can spread very quickly on Twitter because users are often very close to the sources of the news. Users are allowed to receive updates from someone who they can not have connections with in real life, such as a celebrity or politician. However, social connections on Facebook are reciprocal.

Unlike Facebook and Twitter, where users can talk about anything they want, some social networks have particular subjects, e.g., YouTube provides a place that allows people to upload, view and share videos; LinkedIn helps people to build and expand their professional networks and Yahoo!Answer allows people to submit questions to be answered or answer questions asked by others.

This section discusses the history of social networks in the beginning, followed by the mechanisms of social networks and the changes they bring. At the end of this section, the social network used for this study, Sina Weibo, is discussed.

2.1.1 History

SixDegrees.com is the first recognisable social network launched in 1997 (Ellison *et al.*, 2007). Users were allowed to create their own profiles and establish connections with others. It was closed in 2000 even though it had over 3 million registered users at its height (Kirkpatrick, 2010). It did not succeed because it appeared too early. The market just was not ready. People were not accustomed to make friends online. In the next 3 years (2000 - 2003), social networks went through ups and downs. The most notable one was Friendster, launched in 2002. It was originally designed to compete with an online dating website Match.com and 3 million users registered within the first few months (Rivlin, 2006). Because of its extraordinarily fast growth, Friendster experienced technical problems (i.e. its computer system was ill-equipped to handle such large amounts of information). Users suffered from system failures and left the website. During the same time frame, many social networks were created such as Cyworld, Ryze, and LinkedIn. MySpace was created in 2003 and became one of the most successful social networks. Unlike its predecessors, such as Friendster, MySpace keeps adding new features based on users' demand so as to retain old users and attract new users. It even allowed users to personalise their webpages, e.g. embed HTML code in profile pages so as to make them unique and appealing. Facebook was created in 2004 but users were limited to Harvard students. From 2005 to 2006, Facebook began to relax the restriction on users and now it is open to the masses. Facebook allows other developers to embed web-based applications in Facebook so that users can play games with each other, analyse other users' behaviour, compare movie preferences, etc (Ellison *et al.*, 2007). Twitter was created in 2006. It has become the most widely used microblogging site (i.e. a kind of social networks). Twitter differentiates itself by restricting the length of a 'tweet' to 140 characters. Originally, the founders of Twitter thought it would be better if the length of a tweet could be within the bounds of a standard length of Short Message Service (SMS). Although few people view tweets via SMS nowadays, concise tweets are more readable and easier to spread.

2.1.2 Mechanism

Unlike early public discussion forums which were structured by topics, social networks centered around users. A profile is considered as the digital representation of a user's identity. A user profile consist of a set of attributes such as gender, age, education background, profile picture, screen name, interests, self-description, friends list, etc. Usually, it is a window to the public, some users even pretend to be someone else by creating fake profiles. Nearly all forms of social networking websites require users to sign up for them before the first use. Users can provide their profiles at the time of signing up for the website or they can give a little essential information, such as email, and complete the rest later. Profiles can be visible to the public or to specific users only.

Besides profiles (i.e. users themselves), connections among the users, are also indispensable to a social network. In fact, the term 'network' come from the product of the interactions among users. Social connections can be one-way (i.e. asymmetric) or two-way (i.e. symmetric) depending on the particular social network. For example, two-way connections require that both parties (i.e. initiators and recipients) agree to build connections between them; on the other hand, one-way connections allow users to add anyone to their friends lists without approval. Adding friends on Facebook is called 'Add Friend'; adding someone to a circle on Google+ is called 'Add'; subscribing to someone's microbiolog on Twitter is called 'Follow'. Although social networks name it differently,they use either a one-way model or a two-way model. For example, Facebook is a typical example of two-way connections but Twitter uses a one-way model.

A user creates a link to another user for numerous reasons. The target user can be a real-world acquaintance, such as a friend, family member or a business partner. The target user can also be a virtual acquaintance, such as a celebrity or alumni member, that they have never met before. That is why social networks are so addictive.

Users and the connections among them are the skeleton of a social network but the content is the key to attract new users and retain old users. Social networks have similar skeletons but the content can be varied. For example, YouTube allows users to upload, view and share videos but Twitter emphasises a quick, frequent and smart read; Facebook aims to provide an all-in-one platform but Instagram focuses on photo-sharing.

2.1.3 New Forms of Communication

The past decade has witnessed a tremendous change in the forms of communication Web 2.0 has made. Social networking is a typical example of the application of Web 2.0. First, it is easier to make new friends. Before the advent of social networking websites like Facebook and LinkedIn, being friends with someone required physical presence such as working at the same place, going to the same school, ect. Now, people can find new friends through the filters provided by a social networking website. Second, people can interact with anyone anywhere anytime. Most social networks use push technology to forward information to users' mobile devices, whether the users request it or not. Push technology is contrasted with pull technology where a user must request the information before it becomes available. Third, everyone can have a voice. Unlike traditional media such as magazines and newspapers, where only a small number of experts can express their opinions to the public, social networks provide a mean for ordinary people to reach a large audience (e.g. having an Instagram account of thousands of followers does not require being a celebrity). While people certainly benefit from the use of social networks, there are also some downsides such as being addicted to social networks. As the advantages and disadvantages of social networks have been widely discussed (Fuchs, 2013; Fraser and Dutta, 2008), and will not be considered further in this thesis.

2.1.4 Sina Weibo

Sina Weibo (Weibo) is a microblogging social networking service launched on 14 August 2009. Coincidentally, China blocked access to Twitter on 2 June 2009 (Branigan, 2009), 2 months before the launch of Weibo. Although most people consider Weibo a Chinese Twitter, it does have some unique features which differentiate from Twitter, such as threaded comments, rich media, categorised trends, a hall of celebrities, reward systems, etc. Bishop stated "It is not fair to call Sina Weibo a Twitter clone or knockoff. It is a better designed and more stable product, and Sinas product roadmap appears to have it heading towards a robust SNS, almost like Facebook. I hope Twitter has people dissecting Weibo, as they could learn a lot." (Bishop, 2011)

Weibo was ranked 16th in 2014 with respect to its global traffic (Alexa, 2014). Monthly Active Users (MAUs) reached 156.5 million as of 30 June 2014 (Sina, 2014), compared with 271 million MAUs on Twitter at the same time. It is a great achievement considering that Weibo is China-based but Twitter is worldwide. 94.5% of visits come from mainland China, followed by United States at 1.2% and Taiwan at 1.1%. In terms of the frequency of visits, Weibo is the fifth most popular website in China (see Table 2.1). In fact, it is the most popular social network in China. Without doubt, Weibo has achieved huge success in China, even celebrities e.g. Tom Cruise (over 5 million followers on Weibo, compared with 4.7 million followers on Twitter), Bill Gates, Emma Watson and Maria Sharapova who do not speaking Chinese at all use Weibo to extend their influence in China (Moore, 2011).

Table 2.1: Most popular websites in China (Alexa, 2014)

Rank	**Domain**	**Description**
1	baidu.com	the leading search engine in China
2	qq.com	Internet portal owned by Tencent Holdings Limited
3	taobao.com	online shopping
4	sina.com.cn	Internet portal owned by Sina Corp
5	weibo.com	microblogging service provided by Sina Corp
6	hao123.com	Internet portal owned by Baidu
7	tmall.com	online shopping
8	sohu.com	Internet portal owned by Sohu, Inc.
9	360.cn	offers comprehensive Internet and mobile security products and services for free
10	soso.com	search engine in China

2.2 Social Network Analysis

Social Network Analysis (SNA) examines social networks using graph theory. Users are viewed as vertices and interactions between them are represented by edges. The resulting graph makes detection of the patterns of interactions much easier.

SNA has drawn extensive attention from all walks of life. Common applications include network modelling, centrality analysis, community detection, classification, etc. Businesses use it to support activities such as public relations, customer analysis, advertising, etc. Governments use it to detect terrorist threats, develop communities, educate civilians, etc. A wide range of sciences are involved in SNA, such as social science, behaviour science, computer science, etc. Research areas in

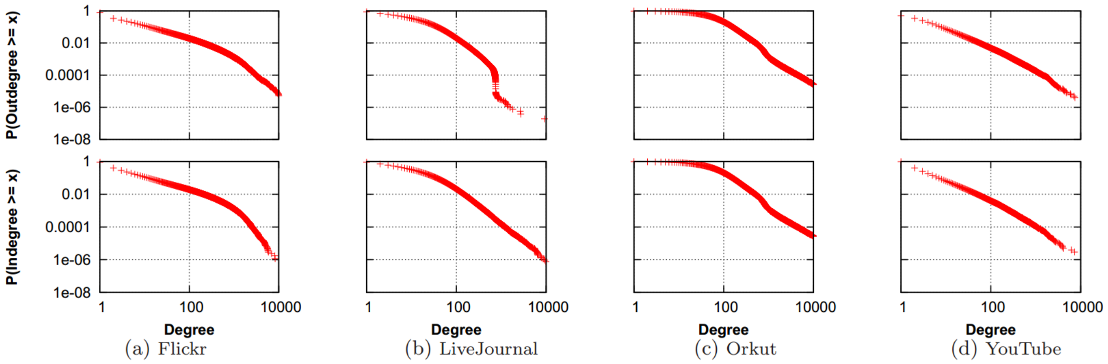
SNA are being expanded as more data becomes available and specialists from other disciplines also pay attention to it.

This section consists of two parts. The first part discusses the properties of social networks. Recognising them helps researchers to be able to make proper decisions on the tasks in relation to SNA, such as data collection, data sampling, and data analysis. The challenges of analysis of the large-scale data from social networks are illustrated in the second part. These challenges can drive research in different directions, which are discussed in Section 2.3.

2.2.1 Properties of Social Networks

Social networks are often huge, such as dozens of millions of users, and the interactions between these users are even more complex. No matter what kinds of social networks they are, they do have some things in common that are seldom seen in other kinds of networks. The most well-known common properties include power law distributions, the small-world effect and strong community structure.

The degrees of vertices in social networks often follow power law distributions or long tail distributions. To be more specific, vertices with lower degrees are more frequent than vertices with higher degrees. Figure 2.1 indicates that power law distributions exist in Flickr, LiveJournal, Orkut, and YouTube.



(a) Flickr (b) LiveJournal (c) Orkut (d) YouTube

Figure 2.1: The proportion of in-degree and out-degree for each type of social networks. Calculation based on complementary cumulative distribution functions. (Mislove *et al.*, 2007)

Another characteristic of social networks is the samll-world effect (or six degrees of separation). Half century ago, scientists (Travers and Milgram, 1969) had already investigated the average path length between people in Nebraska and Boston and people in Massachusetts. The results showed that anyone is just six relationships away from anyone else on Earth. The small-world effect exists in social networks as well. The average path length on Facebook was 4.7 (Ugander *et al.*, 2011), that on Twitter was 4.12 (Kwak *et al.*, 2010). The average path length on YouTube was a little bit longer, at 5.1 (Mislove *et al.*, 2007). This could be participially explained by the fact that Mislove's study was conducted in 2007, when social networking was in its infancy.

Social networks also show a strong community structure. This means that people within the same community tend to interact with each other more frequently. On the other hand, people from different communities barely connected to one another. We use 'clustering coefficient' to measure the degree to which users on a social network tend to cluster together. The clustering coefficient C_i for vertex v_i is defined as follows:

$$C_i = \begin{cases} \frac{2k}{m(m-1)} & m > 1 \\ 0 & m = 0 \text{ or } 1 \end{cases} \tag{2.1}$$

where m is the total number of neighbours of v_i and k is the total number of edges among the neighbours of v_i. A network with communities tends to have a higher average clustering coefficient than a random network. The average clustering coefficient $\bar{C}$ for a network is defined as follows:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{2.2}$$

where n is the total number of vertices on the network. Figure 2.2 shows two graphs, one with a community structure and one without a community structure. In Figure 2.2(a), vertex v_3 has 3 neighbours (i.e. $m = 3$) and 1 mutual connection (i.e. $k = 1$) among the neighbours (i.e. $\langle v_1, v_2 \rangle$), according to Equation 2.1, the clustering coefficient for vertex v_3 is calculated as follows:

$$C_3 = \frac{2 \times 1}{3 \times (3-1)} = \frac{1}{3}$$

According to Equation 2.2, the average clustering coefficient for Figure 2.2(a) is calculated as follows:

$$\bar{C} = \frac{C_1 + C_2 + C_3 + C_4 + C_5 + C_6}{6} = \frac{1 + 1 + \frac{1}{3} + \frac{1}{3} + 1 + 1}{6} = \frac{7}{9}$$

which indicates how likely the vertices in Figure 2.2(a) are to be clustered. In contrast, the average clustering coefficient for Figure 2.2(b) is calculated as follows:

$$\bar{C} = \frac{C_1 + C_2 + C_3 + C_4 + C_5 + C_6}{6} = \frac{0 + 0 + 0 + 0 + 0 + 0}{6} = 0$$

where $m = 1$ for C_i in Figure 2.2(b). In other words, it is very unlikely for the vertices in Figure 2.2(b) to be clustered.

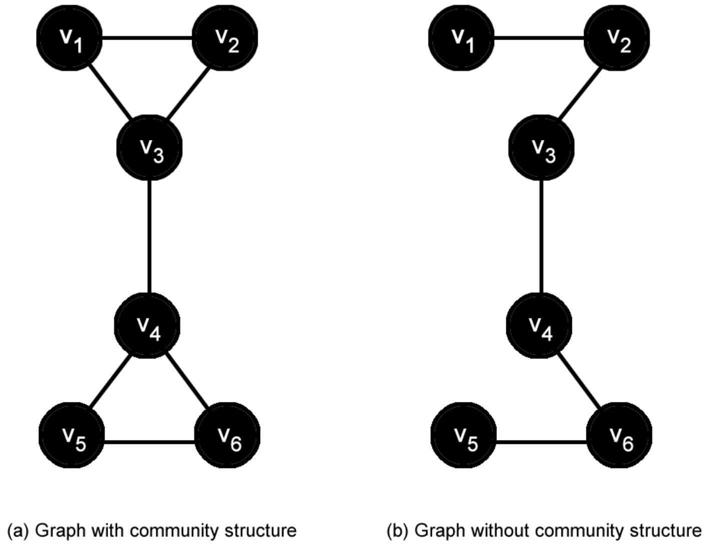

v1
v2
v3
v4
v5
v6
(a) Graph with community structure
v1
v2
v3
v4
v5
v6
(b) Graph without community structure

Figure 2.2: Comparison of average clustering coefficient between two graphs

2.2.2 Challenges

Millions of people play online, learn online and even work online. People are living in an information explosion era, they have never experienced before. Almost everything on Earth has a digital footprint on the Internet. The full capability of SNA has yet to be reached. Social networks combined with their unique characteristics pose challenges that have never been met before. This subsection discusses the challenges of SNA from the following perspectives:

Scalability Millions of users communicate to each other everyday in social networks. The interactions among them are quite complex. Traditional methods for SNA were designed for the analysis of a network that usually consisted of hundreds of entities, e.g. sociologists often carry out a survey on a scale of hundreds or thousands of people to analyse human behaviour. The methods used by sociologists do not work properly when dealing with huge volumes of multidimensional data on the Internet.

Heterogeneity The forms of the interactions can vary, even different forms of interactions exist between the same set of users, e.g. two users work at the same company but they do not like each other. Multiple types of entities are also involved, e.g. user A works together with user B, but connected to user B through cloud sever S. Analysis of these heterogeneous networks requires new theories and models.

Evolution Time matters for SNA. People are likely to lose their attention very quickly. New users join in, old users leave every day and social connections change every day. Uncovering dynamics of social networks is a big challenge.

Collective Intelligence People share their thoughts online in the form of comments, reviews, ratings, etc. Such meta information is useful for many applications. Collecting the intelligence from such data effectively is not a straightforward job but it is very necessary because this intelligence is very precious.

Evaluation Traditional data mining techniques use a training set as benchmark data. However, only a small amount of the data available is suitable for being a training set for two reasons. First, some sensitive information is protected by the privacy policies. Second, it is nearly impossible to get an accurate benchmark because of the dynamic and large-scale nature of social networks.

2.3 Recent Research

Some research on the challenges associated with SNA has already been conducted. Some of them are illustrated here with examples, including network modelling, centrality analysis, community detection, classification and recommendation, and ethics, privacy, security and spam.

2.3.1 Network Modelling

Since the seminal work by Watts and Strogatz (Watts and Strogatz, 1998), and Barabási and Albert (Barabási and Albert, 1999), remarkable progress has been made in network modeling (Chakrabarti and Faloutsos, 2006). Scientists have found that large-scale networks, no matter how they are presented, follow similar patterns such as a power-law distribution, clear community structure and six degrees of separation. Given these patterns, it is intriguing to model the network dynamics of repeated patterns with some simple mechanics. Examples include the Watts-and-Strogatz model (Watts and Strogatz, 1998), to explain the small-world effect, and the preferential attachment process (Barabási and Albert, 1999) that explains power-law distributions. Network modelling offers an in-depth understanding of network dynamics that is independent of network domains. A network model can be used for the simulation study of various network properties, e.g., robustness of a network under attack, or information diffusion within a given network structure, etc.

Intensive computing is required when conducting the analysis of a network consisting of millions of vertices. Sampling is necessary when the computation is too expensive in terms of the computational time and the memory usage required (Leskovec and Faloutsos, 2006). Sampling provides an approximation of different statistics by investigating a small portion of the original huge network. An alternative approach of dealing with large-scale networks is to improve the efficiency and effectiveness of the computation (Becchetti *et al.*, 2008; Desikan and Srivastava, 2008).

2.3.2 Centrality Analysis

Centrality analysis is about identifying critical vertices in large-scale networks (Wasserman and Faust, 1994). Traditional SNA hinges upon link structure to identify vertices with high centrality. Commonly used criteria include: degree centrality, betweenness centrality, closeness centrality and eigenvector centrality (or Pagerank scores). With the rapid increase in the amount of information available for SNA, e.g., users are linked by common interests, the study of centrality in social networks is no longer limited to the structure of the graphs, a user, who initiated a topic, is identified as critical even though that user is not at the centre of the network (Agarwal *et al.*, 2008).

A related task of the study of centrality is to determine how things on a large-scale network influence each other, e.g., the way people affect one another in social networks. Researchers have put considerable effort into identifying the critical users who can help to improve the spread of information (Kempe *et al.*, 2003). An important application in the marketing domain is viral marketing (Richardson and Domingos, 2002), which aims to achieve the maximum return by identifying influential customers for marketing so that they can effectively influence their friends.

2.3.3 Community Detection

Communities are named differently in different fields, such as groups, clusters, etc. Identifying the communities on a social network is a fundamental problem for SNA. Actually, sociologists believe that studying the characteristics of a community as a whole is more meaningful than exploring the individuals' characteristics in the context of social science (Hechter, 1988). Identifying a community in a social network is equivalent to identifying a set of vertices that are densely connected within the group but sparsely connected with the rest. For instance, Figure 2.3 shows the effect after the vertices in (a) are grouped into two different sets in (b) based on modularity optimisation (Newman and Girvan, 2004), with each group in a different colour.

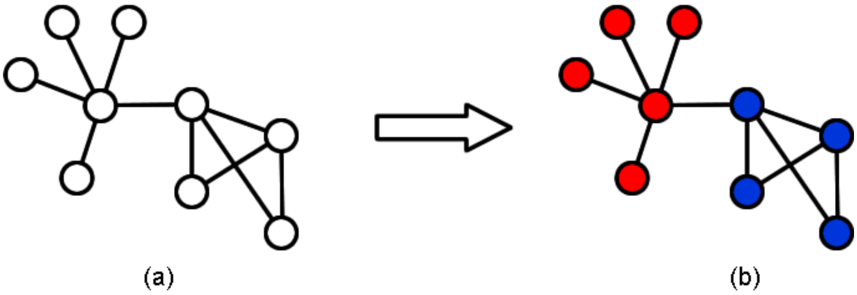


Figure 2.3: Two communities identified by a modularity optimisation method

Community detection can facilitate other social computing tasks and is applied in many real-world applications. For instance, the grouping of customers with similar interests in social media renders efficient recommendations that expose customers to a wide range of relevant ties to enhance transaction success rates. Communities can also be used to compress a huge network, resulting in a smaller network. In other words, problem solving is accomplished at community level, instead of vertex level. In the same spirit, a huge network can be visualised at different resolutions, offering an intuitive solution for network analysis and navigation.

The fast expansion of social networks has spawned novel lines of research on community detection. The first line focuses on making the methods for community detection more suitable for large-scale networks (Flake *et al.*, 2000; Gibson *et al.*, 2005; Dourisboure *et al.*, 2007; Andersen and Lang, 2006). This is because many well-studied approaches in social science were not designed to handle the volume of data in social networks.

The second line of research emphasises the heterogeneous nature of social networks (Zeng *et al.*, 2002; Java *et al.*, 2008; Tang *et al.*, 2008, 2009). For example, individual users and video content are encompassed in YouTube. Different types of entities can interact with each other. The interactions can be sending an emoji, giving a thumbs-up, reposting others' content, etc. These types of interactions form a heterogeneous network, that allows researchers to investigate how communities in one type correlate with those in another type and to determine the hidden communities among heterogeneous interactions.

The third line of research integrates the time dimension with SNA. Because of the dynamic nature of a social network, the members of a community always change, e.g., the number of active users in Facebook increased from 14 million in 2005 to 500 million in 2010. Exploring the evolution of the communities on a social network is important as well (Backstrom *et al.*, 2006; Palla *et al.*, 2007; Asur *et al.*, 2009; Tang *et al.*, 2012).

2.3.4 Classification and Recommendation

Many social networks have recommendation systems. For example, Weibo will suggest the users that are likely to interest you, YouTube will recommend the videos that are similar to what you are viewing now, etc. A good recommendation system is a secret weapon to encourage users to remain active.

Most of recommendation systems are based on classification models. For example, recommending new movies to the users on a social network requires the identification of the cinephiles first, as it is meaningless to recommend a movie to a user who does not like watching movies. Another example is to suggest someone that a user is likely to become friends with. This problem is known as link prediction (Liben-Nowell and Kleinberg, 2007). Basically, it is about predicting which pairs of vertices are likely to be connected to each other.

Figure 2.4 is an example of link prediction. The starting point is the network on the left side. Based on the network structure, link prediction generates a list of connections that are most likely. In the example, one connection is suggested: $\langle v_1, v_2 \rangle$, resulting in the network on the right in which the dashed line is the predicted link. If a network involves more than one type of entity, the recommendation becomes a collaborative filtering problem (Breese *et al.*, 1998).

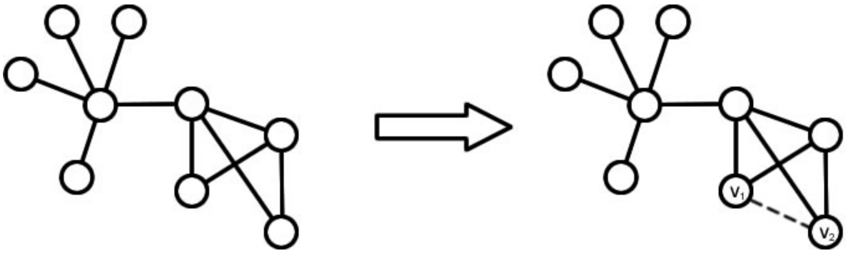


Figure 2.4: Link prediction where dashed line is the predicted link

There are other tasks that also involve the utilisation of social networks. For example, inferring the missing part of a user's profile based on other users that have similar profiles. Figure 2.5 presents an example of network-based classification, where smokers are coloured in green with "+", non-smokers are coloured in red with "−", and users with unknown smoking behaviour are coloured in yellow with "?". By studying connections between users, it is possible to infer the behaviour of those unknown users as shown on the right. Social networks offer rich information to researchers for the study of human behaviour.

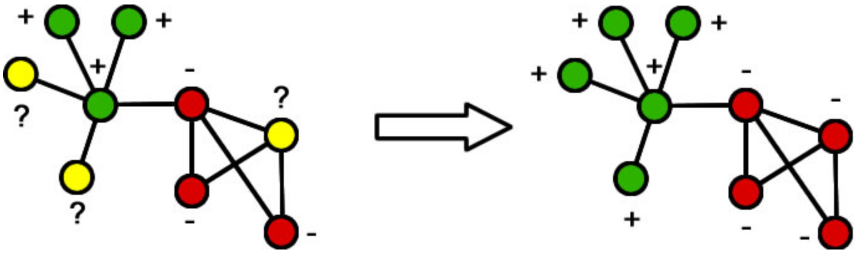


Figure 2.5: Network-based classification

2.3.5 Ethics, Privacy, Security and Spam

Ethics, privacy and security are a set of inevitable and sensitive topics. Discussions about the ethics, privacy and security of social networks are often heated and controversial. Issues of ethics, privacy, and security can not be dealt with lightly. For example, Facebook is facing criticism of manipulating the public agenda (BBC, 2014). Spam is another issue that has attracted a great deal of attention. Spam includes, but is not limited to, malvertisements, phishing scams, malicious links and other unwanted content. Preventing people from being affected by spam is more and more important.

Chapter 3

Mathematical Theories Behind Social Network Analysis

This chapter reviews the theories that are relevant to SNAHOC. First, graph theory with respect to SNA is discussed. Then, the algorithm used to collect data from social networks is given. As SNAHOC is built upon a hierarchy of communities, the methods used to identify communities in social networks are presented. The remaining part of this chapter delineates the techniques used in case studies.

3.1 Graph Theory in Social Networks

Graph theory is an old branch of mathematics. Nowadays, it is an important tool in computer science, sociology, chemistry, physics, biology, etc. Graph theory has the ability to formulate things easily and precisely, e.g. representing a social network as a graph. This section will introduce a few concepts from graph theory that have been widely used to analyse social structures.

Definition 3.1.1. A graph, G, consists of a set of vertices (or nodes, points) V and a set of edges (or lines, arcs) E. Edges can be directed or undirected. In a directed graph, $\langle v_1, v_2 \rangle$ and $\langle v_2, v_1 \rangle$, where $v_1, v_2 \in V$ represent different edges. In an undirected graph, $\langle v_1, v_2 \rangle$ and $\langle v_2, v_1 \rangle$ represent the same edge.

Definition 3.1.2. If $G = (V, E)$ is undirected, v_1 is adjacent to v_2 if $e = \langle v_1, v_2 \rangle$ exists, where $v_1, v_2 \in V$ and $e \in E$. Both v_1 and v_2 are two endpoints of e_1.

Definition 3.1.3. If $G = (V, E)$ is undirected, $e_1 = \langle v_1, v_2 \rangle$ is adjacent to $e_2 = \langle v_1, v_3 \rangle$ because both e_1 and e_2 have the same endpoint v_1.

Definition 3.1.4. A graph that does not have multiple edges (where two vertices are connected by more than one edge) is called a simple graph.

Definition 3.1.5. A complete graph is a simple graph where each vertex is adjacent to every other vertex.

Definition 3.1.6. A path in a graph represents a way to get from an origin (a vertex) to a destination (another vertex) by traversing edges in the graph.

Definition 3.1.7. A graph is connected when there is a path between every pair of vertices.

Definition 3.1.8. A graph $G' = (V', E')$ is a subgraph of $G = (V, E)$ if $V' \in V$ and $E' \in E$.

Definition 3.1.9. A connected component of an undirected graph is a subgraph in which any two vertices are connected to each other by paths. A directed graph is called weakly connected if replacing all of its directed edges with undirected edges produces a connected graph. It is strongly connected if it contains a directed path from v_i to v_j and a directed path from v_j to v_i for every pair of vertices v_i and v_j.

Definition 3.1.10. The degree of a vertex of a graph is the number of edges incident to the vertex. The in-degree of v is the number of edges with v as the terminating vertex. The out-degree of v is the number of edges with v as the initiating vertex.

3.2 Breadth-first Search

Breadth-First Search (BFS) starts at a given vertex, s. Vertices at a distance of 1 from s (i.e. adjacent to the vertex, s) are discovered. Then, vertices adjacent to the vertices at a distance of 1 from s are discovered, which are at a distance of 2 from

s. BFS keeps exploring the graph in this particular way until every vertex that is reachable from s is discovered (Cormen *et al.*, 2001).

```
Breadth-first Search($V$, $E$, $s$)
for each $u \in V$ do
  colour[$u$] $\leftarrow$ WHITE
end for
colour[$s$] $\leftarrow$ GREY
$Q \leftarrow \emptyset$
ENQUEUE($Q$,$s$)
while $Q \neq \emptyset$ do
  $u \leftarrow$ DEQUEUE($Q$)
  for each $v \in$ Adj[$u$] do
    if color[$v$]=WHITE then
      colour[$v$] $\leftarrow$ GREY
      ENQUEUE($Q$, $v$)
    end if
  end for
  colour[$u$] $\leftarrow$ BLACK
end while
```

The procedure BFS is given above. The colour of each vertex $u \in V$ is stored in colour[u]. Q is a first-in, first-out queue. Adj[u] represents the vertices at distance 1 from u.

3.3 Community Detection

Numerous community detection methods have been introduced over the past few years. According to a comparative study by Santo (Fortunato, 2010), methods by Rosvall and Bergstrom (Rosvall and Bergstrom, 2007, 2008), Blondel et al. (Blondel *et al.*, 2008) and Ronhovde and Nussinov (Ronhovde and Nussinov, 2009) have better performance than the others. This section mainly focuses on the first two methods.

3.3.1 Multilevel Community Detection

Blondel et al. (Blondel *et al.*, 2008) presented a simple method based on the optimisation of the modularity (see Equation 3.1) to identify the community structure. It is very fast in terms of computational time (i.e. $O(m)$) although memory requirements grow quickly as well. Their results show that the algorithm they proposed is capable of handling a complex network with 100 million vertices.

Blondel et al.'s method is made of two phases: identifying the community structure of a network by optimising the modularity of the network and creating a new network of communities by contracting the communities found in the previous phase into vertices of the new network. These two phases are repeated iteratively until the modularity of the network can not be improved.

In the first phase, each vertex is assigned to a different community. There are as many communities as there are vertices. The modularity Q is calculated as follows (Blondel *et al.*, 2008):

$$Q = \frac{1}{2m} \sum_{v_i, v_j} \left[w(v_i, v_j) - \frac{k_{v_1} k_{v_j}}{2m} \right] \delta(c_{v_i}, c_{v_j}) \tag{3.1}$$

where $w(v_i, v_j)$ represents the weight of $\langle v_i, v_j \rangle$, $k_{v_i} = \sum w(v_i)$ is the sum of the weights of the edges incident to v_i, $k_{v_j} = \sum w(v_j)$ is the sum of the weights of the edges incident to v_j, c_{v_i} is the community v_i belongs to, c_{v_j} is the community v_j belongs to, m is the sum of the weights of the edges in the given network and $\delta(c_{v_i}, c_{v_j})$ is 1 if $c_{v_i} = c_{v_j}$ and 0 otherwise.

For each vertex v_i, the increase in modularity obtained by moving v_i from its current community to the communities that are adjacent to it, is evaluated. The increase in modularity, ΔQ is defined as follows (Blondel *et al.*, 2008):

$$\begin{aligned}\Delta Q = \left[\frac{\sum_{in} + k'_{v_i}}{2m} - \left(\frac{\sum_{out} + k_{v_i}}{2m}\right)^2\right] - \\ \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{out}}{2m}\right)^2 - \left(\frac{k_{v_i}}{2m}\right)^2\right]\end{aligned} \tag{3.2}$$

where $\sum_{in}$ is the sum of the weights of the edges inside the community, $\sum_{out}$ is the sum of the weights of the edges incident to the vertices of the community, k_{v_i} is the sum of the weights of the edges incident to vertex v_i no matter whether the other endpoint is inside the community or not, k'_{v_i} is the sum of the weights of the edges incident to vertex v_i and the other endpoint must be inside the community as well, and m is the sum of the weights of the edges in the given network. The vertex, v_i, is then put into the community for which ΔQ reaches the maximum. If no increase is possible, v_i stays in its original community. This process is repeated for all the vertices in the network until no further improvement can be achieved.

In the second phase, communities found in the previous phase are replaced by vertices. Two vertices are connected if at least one edge exists between the communities the vertices represent. The weight of the edge between vertices is the sum of the weights of the edges between the communities the vertices represent. A new network is then formed where the vertices of the network represent the communities found in the previous network.

It is worth noting that the calculation of the modularity (globally) is always based on the original network but the increase in modularity (locally) is calculated based on the interim networks. Blondel et al.'s method identifies the community structure of a network by optimising the increase in modularity, ΔQ, at multiple levels until the modularity of the original network, Q, can not be further improved.

3.3.2 InfoMap

Rosvall and Bergstrom (Rosvall and Bergstrom, 2007, 2008) gave a new perspective on identifying community structures in directed and weighted networks. Community detection is turned to a communication process (Shannon, 2001), in which a complex network is compressed (or encoded) such that the most information about the network can be decoded later.

Given a network X, Y is a simpler description that summarises the structure of X. The best description is the one that tells the most about X whilst unimportant details are filtered out. The information about X that Y does not cover is defined as below (Rosvall and Bergstrom, 2007, 2008):

$$H(X|Y) = log \left[\prod_{i=1}^{q} \begin{pmatrix} n_i(n_i-1)/2 \\ l_i \end{pmatrix} \prod_{i>j} \begin{pmatrix} n_i n_j \\ l_{ij} \end{pmatrix} \right] \tag{3.3}$$

where q is the number of communities, n_i is the number of vertices in community i, l_i is the sum of the weights of the edges inside community i, l_{ij} is the sum of the weights of the edges between community i and j. $H(X|Y)$ reaches the minimum (or $H(X|Y) = 0$) when $X = Y$. In this case, although Y tells everything about X, Y is also too big to be accepted. Thus, InfoMap (i.e. the name of Rosvall and Bergstrom's algorithm) uses the minimum description length principle (Rissanen, 1978; Grünwald *et al.*, 2005) at the same time so as to achieve a better trade-off between a good compression and enough information about the original network.

This method is capable of handling a network with up to 10 thousands vertices (Fortunato, 2010). Unlike the method (i.e. Blondel et al.'s method) mentioned in Section 3.3, InfoMap can be applied to directed graphs.

3.4 PageRank

PageRank was originally designed for ranking websites on Google's search engine. Nowadays, it is widely used to measure a user's influence in social networks.

PageRank is Google's patented algorithm to examine the entire link strcuture of the web and determine which pages are most important. The details of the algorithm are secret, but the main ideas are well-known and much-copied. The general idea about PageRank is described as follows (Page *et al.*, 1999):

$$R(i) = \sum_{j \in B(i)} \frac{R(j)}{N(j)}. \tag{3.4}$$

where $R(i)$ is the PageRank score (i.e. PageRank) of the page, i, $B(i)$ represents all of the pages that link to i and $N(j)$ is the number of outbound links going out from j. A page that is linked to by many pages which have high PageRank scores is likely to have a high PageRank as well.

Equation 3.4 assumes that all of the page links have the same weight; however, this assumption is not always suitable. A pulp magazine, for example, may sell very well, and which is likely to have a high PageRank score according to Equation 3.4, but may not be highly valued. Thus, links are weighted and the PageRank scores are distributed based on the importance of the pages (Haveliwala, 2003; Yu *et al.*, 2004; Ding *et al.*, 2009).

In SNA, Equation 3.4 is interpreted in a slightly different way, where $R(i)$ is the PageRank score of the user, i, $B(i)$ represents all of the users that are following i and $N(j)$ is the number of followers j has. In this thesis, a modified PageRank algorithm has been developed and is used to measure the influence of users in social networks.

3.5 Shannon-Wiener Index

Shannon-Wiener Index (SWI) (see Equation 3.5) was originally designed for measuring the species diversity of a community. It is determined by both the number of species within a community and how evenly they are distributed within the community. It is defined as follows (Molles and Cahill, 1999).

$$H' = -\sum_{i=1}^{s} p_i log_e p_i. \tag{3.5}$$

where H' is the value of the Shannon-Wiener diversity index; p_i is the proportion of the i-th species, and log_e is the natural logarithm of p_i and s is the number of species in the community. The minimum value of H' is 0, which is the value of H' for a community with a single species.

The evenness of a community is defined as follows (Mulder *et al.*, 2004).

$$E = \frac{H'}{log_e s} \tag{3.6}$$

where J' denotes the evenness of a community; H' is the value of the Shannon-Wiener diversity index; s is the number of species in the community and log_e is the natural logarithm of s.

In this thesis, the SWI is used to measure the geographical diversity of communities.

Chapter 4

Hierarchical Model for Social Network Analysis

In SNAHOC, individual users are grouped into communities. The social connections between the communities generate a new network, that is considered an abstraction of the previous network. SNAHOC defines a hierarchy of communities, in which each level is an abstraction of the previous one. SNAHOC models social networks at multiple levels, so that the analysis can be conducted at different levels of abstraction.

This chapter consists of two parts. In the first part, the basic concepts of SNAHOC are given. Most of them are borrowed from other disciplines such as software engineering and database management but are used differently from their usual environments. In the second part, SNAHOC is explained by using graph theory, where the concepts mentioned in the first part are transformed into mathematical notations.

4.1 Basic Concepts

The concepts are presented in order of the sphere of influence they have. First and foremost, SNAHOC follows a pipeline pattern. Secondly, SNAHOC is of hierarchical structure. Each level is an abstraction of the previous one. Thirdly, information is aggregated every time a network is passed from a lower level of abstraction to a higher level of abstraction.

4.1.1 A Pipeline Design

In software engineering, pipelining is a well-known design pattern, in which a complex problem is decomposed into a sequence of steps. First, pipelining emphasises the sequence of pipes. The output of one pipe is the input of the next. Second, each pipe is an independent entity. The internal workings of each pipe are unseen to the others.

A pipeline is composed of a set of pipes and filters, as shown in Figure 4.1. Although the internal workings of a pipe are a black box to the others, an interface between it and its successor (or predecessor) is explicit to each other. An interface is a shared zone where two adjacent pipes can exchange data with each another. In other words, the input and output must be well defined and the neighbours must be informed about the interface they share.

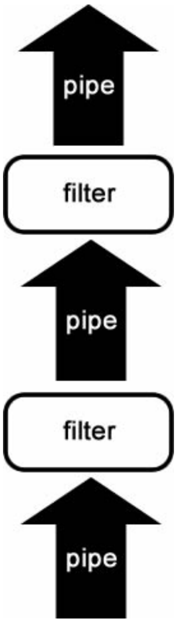


Figure 4.1: A pipeline design

In most cases, the problems SNA deals with can not be solved at one stroke. A complex problem is decomposed into a set of simpler sub-problems. For example, building a recommender system in a social network requires the information of users preferences, items acceptance, the influence from social friends, etc. Thus, the principle of high in cohesion and low in coupling is required to derive a versatile model for SNA, in which sub-tasks can be done independently meanwhile the output of one task can become the input of another task.

SNAHOC adopts the exact same principles as pipelining. A notable feature of using a pipeline pattern is that the internal workings of each pipe are loosely coupled to each other. As a consequence, changing one pipe causes little or no effect on the others.

4.1.2 Zooming

In filmmaking, the term zooming usually refers to the technique of getting a closer view of a far-away object or a wide shot of an object that shows its relationships to its surroundings.

As mentioned in Section 2.2.1, social networks are often huge, dozens of millions of users are involved in. It is necessary to have a mechanism to stand back and look at the big picture rather than the details. For example, in SNA, examining the common behaviour in groups is more useful than exploring the behaviour of a single user. Thus, SNAHOC adopts the concept of zooming in filmmaking for a wide shot. In SNAHOC, the word 'zooming' is used to describe an upward movement on the hierarchy or a downward movement on the hierarchy.

For example, given a graph of 5 vertices, as shown in Figure 4.2. There are 5 vertices at level 3. Each vertex at level 3 represent a community at level 2. When zooming in on the vertex, a, from level 3 to level 2, more vertices are rendered.

It is worth mentioning that zooming in and out in SNAHOC is similar but not identical to zoom in and out on Google map (Google, 2015). A map is of a hierarchical nature. For example, Brooklyn is a district of New York City in the United States of America. Zooming in from a city to a district causes us to descend to a lower level of the hierarchy. As social networks are of hierarchical nature, the

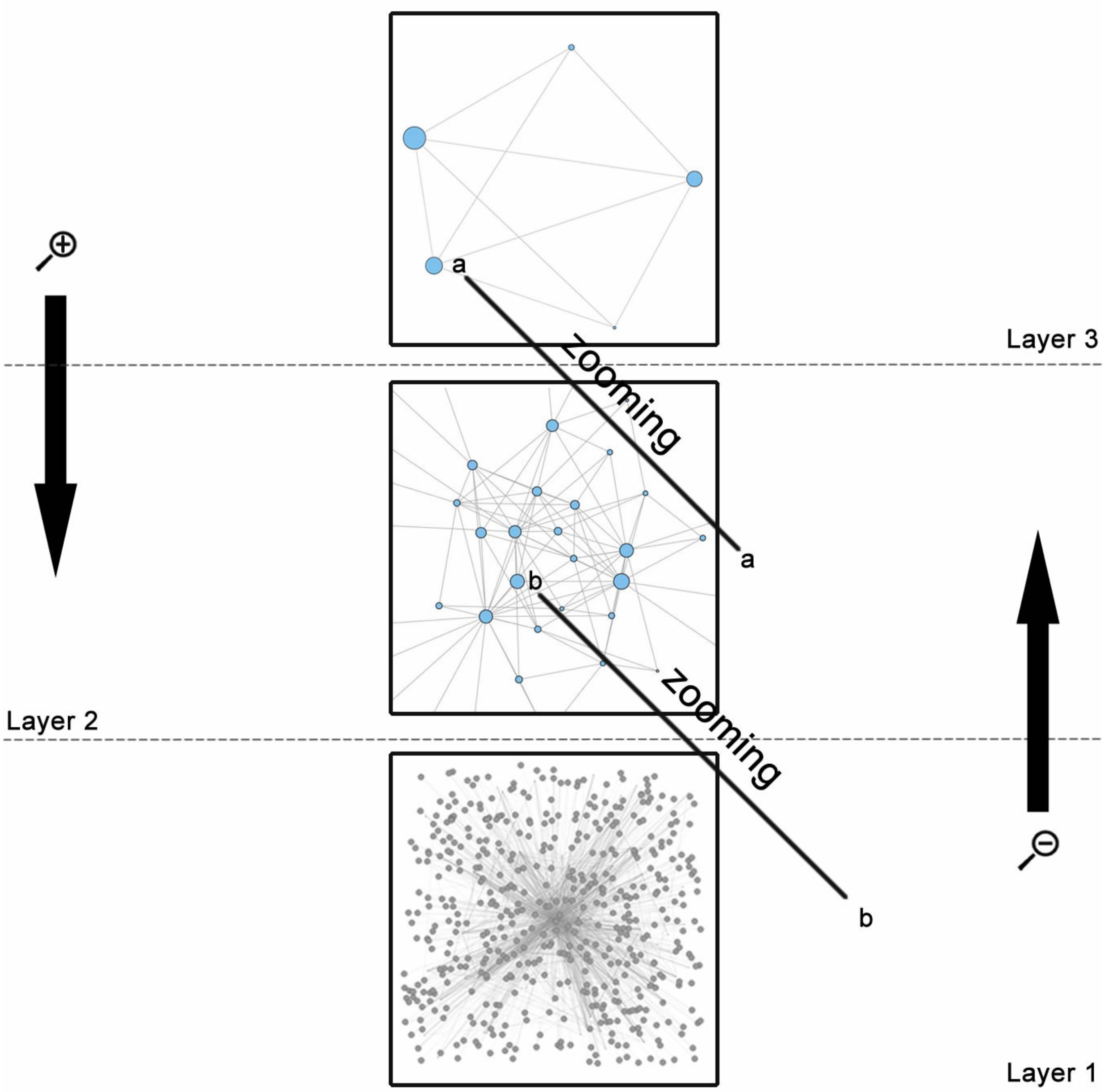


Figure 4.2: Vertices viewed at different levels of abstraction

graphs used to represent social networks are also of hierarchies. The changes of levels of abstraction in SNAHOC follow the hierarchies of the graphs but different levels of abstraction are related indirectly (e.g. the vertex, a, does not exist at level 2, instead, a set of vertices and edges at level 2 are joined together to form a), compared to Google map where different scales are connected directly in a top-down fashion (e.g. a city will never disappear even if you zoom in to a lower level of the hierarchy, like streets). Hierarchies in SNAHOC are established in a bottom-up manner. In other words, level 2 does not exist unless zooming out from level 1. Thus, the input of the model at the very beginning must be the graph at the lowest level of abstraction. The mechanism mentioned above provides the ability to analyse a social network at both a detailed level and an aggregate level.

4.1.3 A Hierarchy of Communities

As social networks have a strong community structure (see Section 2.2.1), in which smaller communities (e.g. a community of one single user) are sparsely embedded in larger communities, such property of social networks can be used to implement a hierarchy of communities. In SNAHOC, vertices are grouped into the same community if they are tightly connected to each other and loosely coupled to the vertices in other communities. A hierarchy of communities can be modelled as a rooted tree, the root of the tree forms the highest level of abstraction and the leaves of the tree form the lowest level of abstraction.

In Figure 4.3, for example, there are 21 vertices at the beginning, which are grouped into 6 communities (i.e. $C_1, C_2, \ldots, C_6$) at the next higher level. Then, Community $C_1, C_2, \ldots, C_6$ are further grouped into 2 communities (i.e. $C'_1 and C'_2$) at the next higher level. Finally, Community C'_1 and C'_2 are grouped into 1 community (i.e. C''_1) at the root level. Figure 4.4 is an equivalent tree structure of the clusters.

A hierarchy of communities is the cornerstone of the 'zooming' operation of SNAHOC.

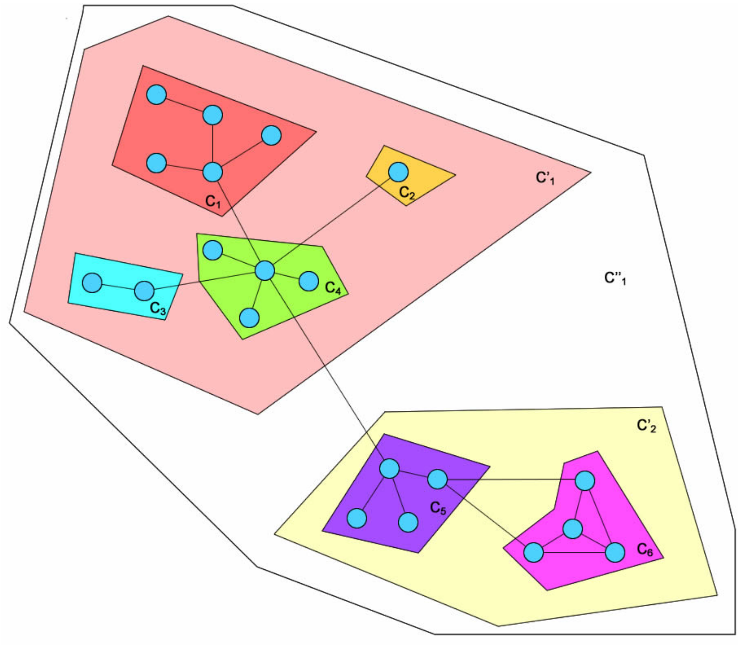


Figure 4.3: A hierarchy of clusters

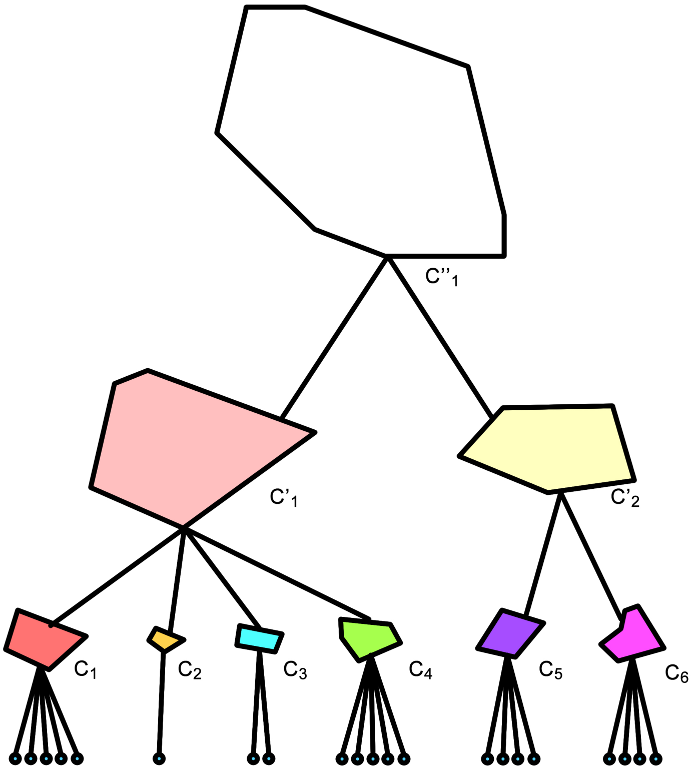


Figure 4.4: An equivalent tree structure of Figure 4.3

4.1.4 Aggregation

In database management, an aggregator is a function of grouping multiple values into a single value, such as sum, count, maximum, minimum, average, etc.

In SNAHOC, not only users are organised in a hierarchy but also the information associated with the users is organised in a hierarchy. Information at one level is always aggregated at the next higher level. As a consequence, valuable information, which is not obtainable at a micro level, can be mined at a macro level.

In Table 4.1, for example, a, b, c and d are the 4 vertices (or users) in Cluster C_4 (or Community C_4) (see Figure 4.3). Information can be aggregated in C_4, for example, the demographic distribution of residence location (75% of New York and 25% of Los Angeles), the majority of interests (75% of users love sports), even complicated aggregation like association rules (e.g. the occurrence of a user who lives in New York and also likes sports is 100%). The information above can be further aggregated in Cluster C'_1.

Table 4.1: Users in Community C_1

Attribute **User**	**Location**	**Interests/Tags**
a	New York	IT, sports
b	New York	fashion, sports
c	New York	sports, travels
d	Los Angeles	IT, reading

This research work uses two case studies to demonstrate how the information is aggregated in the proposed model and how it helps us to better understand a social network at different levels of abstraction.

4.2 Modelling

In this section, a social network is modelled to a hierarchical structure where the social network is divided into discrete levels. Each level represents an abstraction of the social network but with different levels of detail.

4.2.1 User Attributes

Let $G = (V, E)$ be an undirected graph of n vertices and m edges, where vertices are denoted by $V = \{v_1, v_2, \ldots, v_n\}$ and edges are denoted by $E = \{e_1, e_2 \ldots, e_m\}$. Let $D = \{d_1, d_2, \ldots, d_k\}$ be a set of directions and A_{v_i} be a set of vectors attached to v_i where $v_i \in V$ and $1 \leq i \leq n$. $A_{v_i} = \{A_{v_i}(d_1), A_{v_i}(d_2), \ldots, A_{v_i}(d_k)\}$. $|A_{v_i}(d_j)|$ is the magnitude of $A_{v_i}(d_j)$, where $d_j \in D$ and $1 \leq j \leq k$.

The graph, $G = (V, E)$, with V={Mark, John, Lionel, Joss, Harold, Zoe} and E={(Mark, Lionel), (John, Harold), (John, Lionel), (John, Zoe), (Joss, Harold), (Harold, Zoe)} is drawn below, as an example (see Figure 4.5).

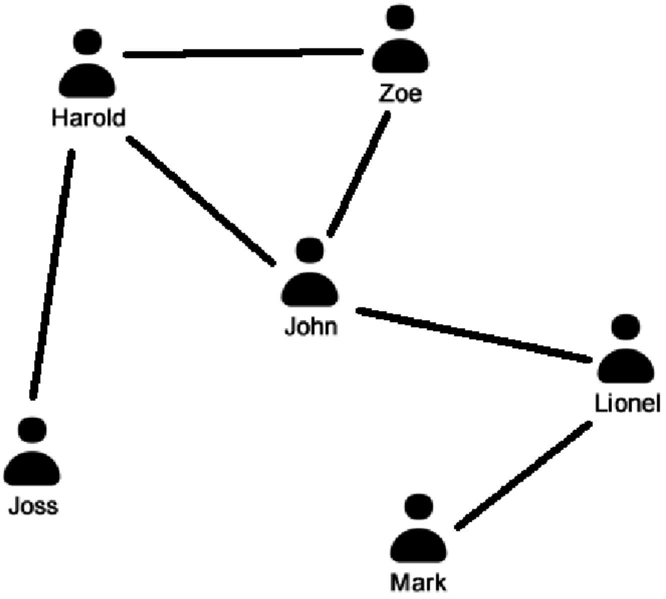


Figure 4.5: An example of a social network

Given D={name, gender, age, location, weibo}, then there are $A_{John}(age) = 35$, $A_{Harold}(location)$=New York, $A_{Joss}(weibo)$=64, etc., as shown in Table 4.2. A_{v_i} is used as a mathematical representation of the profile of the user, v_i. As mentioned above, a user's profile is usually composed of a group of attributes such as name, gender, age, location, etc., thus, d_j is used to refer to one of them and $A_{v_i}(d_j)$ is the value of the attribute of the user.

Table 4.2: An example of vectors

Direction **Vertex**	**Gender**	**Age**	**Location**	**Weibo**
Mark	m	32	Los Angeles	145
John	m	35	New York	223
Lionel	m	41	New York	187
Joss	f	26	Washington, D.C.	64
Harold	m	48	New York	276
Zoe	f	37	New York	129

4.2.2 Hierarchy of Communities

Given an undirected graph $G_1 = (V_1, E_1)$ as shown in Figure 4.6, let P_1=$\{C_{1,1}$, $C_{1,2}$, $\ldots$, $C_{1,k}\}$ be a partition of V_1, such that $C_{1,i} \neq \emptyset$; $C_{1,i} \cap C_{1,j} = \emptyset$; and $C_{1,1} \cup C_{1,2} \cdots \cup C_{1,k} = V_1$.

A subgraph $G_{1,i} = (C_{1,i}, E_{1,i})$ is considered a community of G_1, where $C_{1,i} \subseteq V_1$ and $E_{1,i} = \{(v_x, v_y) \in E_1 | v_x, v_y \in C_{1,i}\}$. Given an undirected graph $G_2 = (V_2, E_2)$ such that each vertex of G_2 represents a community of G_1. Two vertices $v_{2,i}$ and $v_{2,j}$ that correspond to the communities $C_{1,i}$ and $C_{1,j}$ in G_1 are connected in G_2, if and only if there exists $(v_x, v_y) \in E_1$ where $v_x \in C_{1,i}$ and $v_y \in C_{1,j}$. The graph, G_2 as shown in Figure 4.7, is considered more abstract than G_1. Assume given another undirected graph G_3 such that each vertex of G_3 represents a community of G_2, then G_3 is more abstract than G_2. The subscript is used to indicate the level of abstraction (i.e. G_1 represent the network at level 1, G_2 represent the network at level 2, and so forth).

Given a rooted tree T (see Figure 4.8). The leaves of T are denoted by $\mathit{lf}(T)$. The depth of a leaf is denoted by $dp(l_i)$ where $l_i \in \mathit{lf}(T)$. The root of T is denoted by $rt(T)$. The height of T is denoted by $ht(T)$. Then $ht(T) = \max_{1 \leq i \leq n} dp(l_i)$ where n is the number of leaves in T. As $dp(l_i) = dp(l_j)$ for any $l_i, l_j \in \mathit{lf}(T)$, there is $ht(T) = dp(l_i)$ for any $l_i \in \mathit{lf}(T)$.

Given a hierarchy $H = \{G_1, G_2, G_3, \ldots, G_h\}$. The subscripts $1, 2, \ldots, h$ are used to represent the levels of abstraction of a social network. Given a rooted tree T

such that $ht(T) = h$, $V_1 = lf(T)$, and $V_i = V^T_{i-1}$ where $1 \leq i \leq h$. V_i is used to denote the vertices of G_i. V^T_{i-1} is used to denote the vertices of T at height $i - 1$ (i.e., $V^T_{i-1} = \{v \in T | dp(v) = i - 1\}$).

The graph, $G_1 = (V_1, E_1)$, with $V_1 = \{v_{1,1}, v_{1,2}, \ldots, v_{1,9}\}$ and $E = \{(v_{1,1}, v_{1,2}), (v_{1,1}, v_{1,3}), \ldots, (v_{1,8}, v_{1,9})\}$ is shown in Figure 4.6, as an example. Assume $P_1 = \{C_{1,1}, C_{1,2}, C_{1,3}\}$ is a partition of V_1, such that $C_{1,1}$=$\{v_{1,1}$, $v_{1,2}$, $v_{1,3}\}$, $C_{1,2}$=$\{v_{1,4}$, $v_{1,5}$, $v_{1,6}\}$, and $C_{1,3} = \{v_{1,7}, v_{1,8}, v_{1,9}\}$,. Then, $G_{1,1} = (C_{1,1}, E_{1,1})$, $G_{1,2} = (C_{1,2}, E_{1,2})$, and $G_{1,3} = (C_{1,3}, E_{1,3})$ are 3 communities of G_1, such that $E_{1,1} = \{(v_{1,1}, v_{1,2}), (v_{1,1}, v_{1,3}), (v_{1,2}, v_{1,3})\}$, $E_{1,2} = \{(v_{1,4}, v_{1,5}), (v_{1,4}, v_{1,6}), (v_{1,5}, v_{1,6})\}$, and $E_{1,3}$=$\{$ $(v_{1,7}$, $v_{1,8})$, $(v_{1,7}$, $v_{1,9})$,$v_{1,8}$, $v_{1,9})\}$. It is worth mentioning that $(v_{1,3}, v_{1,5})$, $(v_{1,3}, v_{1,7})$, and $(v_{1,5}, v_{1,7})$ do not belong to any of the communities mentioned above as the two endpoints of the edges are in two different groups. Take $(v_{1,3}, v_{1,5})$ as an example, $v_{1,3} \in C_{1,1}$ and $v_{1,7} \in C_{1,3}$.

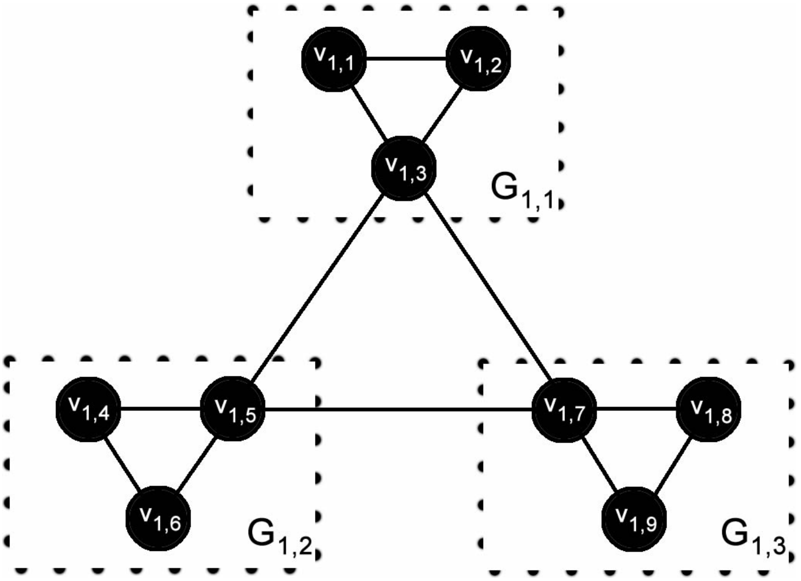


Figure 4.6: Graph G_1 at the minimum level of abstraction

There is $G_2 = (V_2, E_2)$, where $V_2 = \{v_{2,1}, v_{2,2}, v_{2,3}\}$ and $E_2 = \{(v_{2,1}, v_{2,2}), (v_{2,1}, v_{2,3}), (v_{2,2}, v_{2,3})\}$ such that $v_{2,1}$ represents $G_{1,1}$, $v_{2,2}$ represents $G_{1,2}$ and $v_{2,3}$ represents $G_{1,3}$. $(v_{2,1}, v_{2,2}) \in E_2$ because there exists $(v_{1,3}, v_{1,5})$ in G_1 whose endpoints are in $G_{1,1}$ and $G_{1,2}$. And $(v_{2,1}, v_{2,3}), (v_{2,2}, v_{2,3}) \in E_2$ for similar reasons. G_2 is more abstract than G_1.

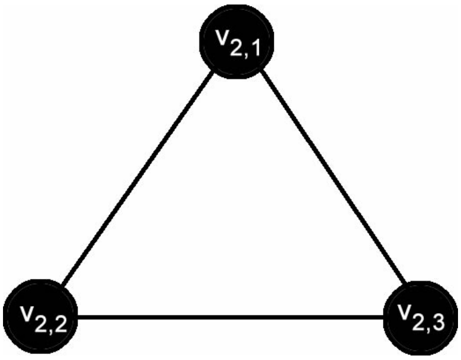


Figure 4.7: Graph G_2 is an abstraction of graph G_1

Assume $P_2 = \{C_{2,1}\}$ is a partition of V_2 such that $C_{2,1} = \{v_{2,1}, v_{2,2}, v_{2,3}\}$. That means all of the vertices in G_2 are in the same group. We have $G_3 = (V_3, E_3)$ with $V_3 = \{v_{3,1}\}$ and $E_3 = \emptyset$ such that $v_{3,1}$ represents G_2. G_3 is more abstract than G_2. There is $G_1 < G_2 < G_3$, which is represented by a tree as shown Figure 4.8.

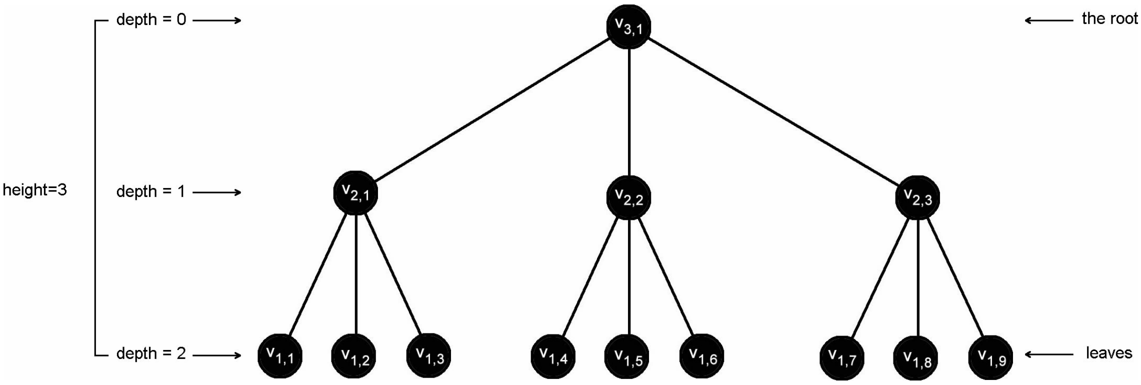


Figure 4.8: T is a rooted tree

Given a rooted tree, T, the vertices of G_1 are the leaves of T, the vertices of G_2 are the vertices of T at height 1 and the vertices of G_3 are the vertices of T at height 2. As shown in Figure 4.8, $ht(T) = 2$ and $dp(v_{1,i}) = 2$ where $i = 1, 2, 3, \ldots, 9$. Figure 4.9 shows a 3D representation of the hierarchical structure $H = \{G_1, G_2, G_3\}$ such that $G_1 < G_2 < G_3$ (see Figure 4.9). It is worth mentioning that there is no direct correspondence between the edges of a graph (i.e. G, G_1 or G_3) and the edges of a tree (i.e. T). T is considered a 2-dimensional representation of H where the connections among the vertices are excluded.

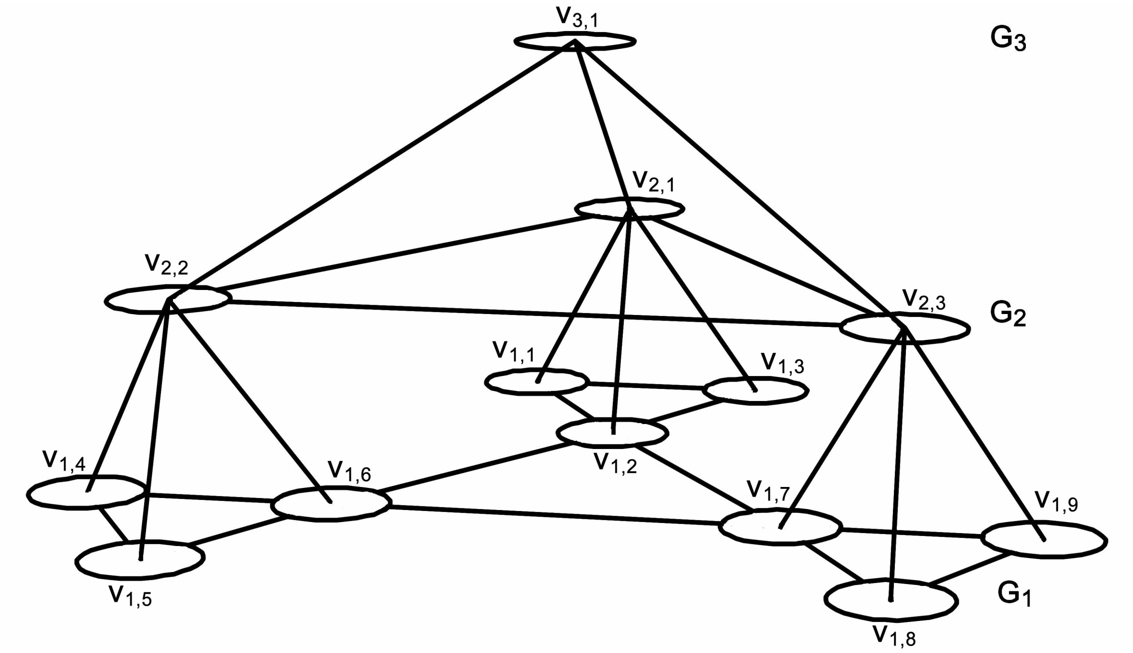


Figure 4.9: A three-dimensional representation of Hierarchy H

4.2.3 Aggregation Function

An aggregation function (i.e. aggregator) is defined as $A_{v_x}(d_i) = f_{agg}(\{A_{v_y}(d_j)|v_y \in C\})$ where: $\{v_y|v_y \in C\}$ is a set of vertices that belong to C and $A_{v_y}(d_j)$ is the vector attached to v_y in the direction of d_j. We use $f_{agg}(\{A_{v_y}(d_j)|v_y \in C\})$ to denote the aggregation of all vertices in C in the direction of d_j. As a result of the aggregation, a new vector of v_x is formed in a new direction d_i. Returning to the example shown in Figure 4.5, assume Mark, John, Lionel, Joss, Harold, and Zoe belong to the same class, say $A_{v_1}(ave\_age)$=$f_{ave}(\{A_{Mark}(age)$, $A_{John}(age)$, $A_{Lionel}(age)$, $A_{Joss}(age)$, $A_{Harold}(age)$, $A_{Zoe}(age)\})$, there exists $A_{v_1}(ave\_age)$=$\frac{32+35+41+26+48+37}{6} = 36.5$. f_{ave} is used as the aggregation function, which returns the average age of the 6 users. Being aware, f_{agg} is not a specific function, it is a general function that can be substituted by different kinds of aggregation functions such as sum, count, maximum, minimum, average, etc.

Aggregation always happens when a graph is passed into a higher level of abstraction (e.g. from G_1 to G_2). One or more aggregation functions are involved. Return to the above example, f_{ave}, f_{max} and f_{min} can be used together so as to find the oldest user in the community, the youngest user in the community, and the average age of the community.

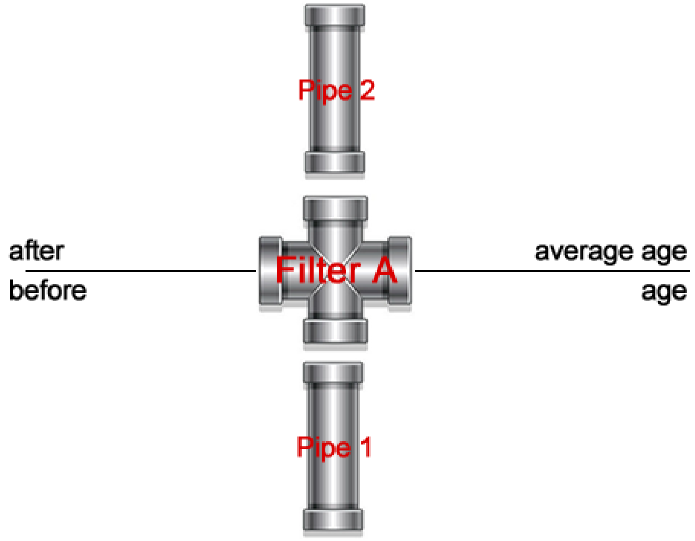


Figure 4.10: Encapsulated functions in Filter A

It is worth mentioning that f_{agg} not only refers to functions like sum, count, maximum, minimum, average, and so forth, but also involves aggregations that are not that intuitive such as finding the most influential user in the community, identifying the most representative tag of the community, etc.

In SNAHOC, aggregation is encapsulated in filters. The output of a filter is considered interim data. The interim data can be the input of the next filter. Assume users' age is the attribute that is aggregated in Filter A (see Figure 4.10). Then, the average age of users in the community is the output of Filter A. Pipeline pattern makes the aggregation much easier to implement.

Chapter 5

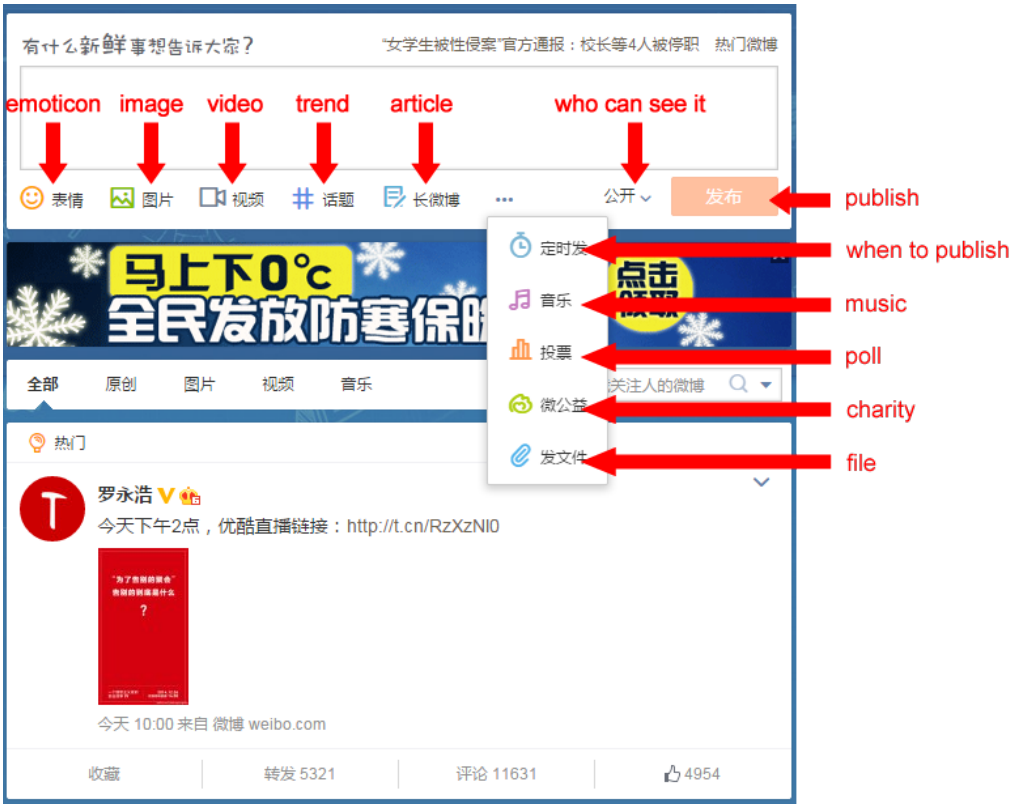
SinaData

SinaData consists of 1,192,972 users and 181,575,370 social connections, retrieved from Sina Weibo. This chapter discusses the data set with respect to data sourcing, data collection and data sampling as well as characteristics of SinaData (Cui *et al.*, 2014).

5.1 Data Sourcing

Sina Weibo was used as a data source, because it is more informative than any other social networks in terms of the contents, the interaction between users and, most importantly, the verification system. Sina Weibo allows users to insert images, videos, music, long articles (more than 140 characters) and even polls without any plug-ins being required (see Figure 5.1). The interaction between users is everywhere, e.g. users are allowed to leave comments on someone's weibo even reply to others' comments on someone else's weibo. Sina Weibo also encourages its users to participate in its identity verification program. Verified users are categorised into 11 groups (see Table 5.1).

Sina Weibo even provides verification services for ordinary people. A 'Pioneer' badge is granted as long as the applicant's real identity is verified and the minimum requirement of being active is satisfied. Unlike 'Pioneer', more information is required so as to grant a badge of one of the rest, e.g. applying for a corporate



有什么新鲜事想告诉大家？

“女学生被性侵案”官方通报：校长等4人被停职 热门微博

罗永浩

今天下午2点，优酷直播链接：http://t.cn/RzXzNI0

今天 10:00 来自 微博 weibo.com

Figure 5.1: The contents on Sina Weibo are more than text

account requires a business license, an official letter with stamp and signature of official representatives, the certificate of trademark registration, the brand letter of authorisation, etc. Because of the strict verification policy, the public has a chance to communicate with the real celebrities and real giants from all walks of life. The trustworthiness and authenticity of the contents posted by verified accounts is guaranteed. Consequently, authenticity stimulates more users to participate actively in Sina Weibo (Chen and She, 2012).

5.2 Data Gathering

Having a complete dataset is a nearly impossible mission for 3 reasons. First, there are more than 500 million users registered on Sina Weibo. Second, privacy is a serious matter for Sina Weibo. Third, only limited access is provided for third-party developers. Instead, a partial dataset is used. To do so, we need to define a termination criterion, choose a search method and specify a type of social relations

Table 5.1: Verified types

Verified type	Domain
Agency	usually referring to welfare organisations, sports clubs, arenas, and other non-governmental organisations
Application Software	usually used to promote the use of an application
Brand	corporate accounts, usually using Sina Weibo to promote their brand values
Campus	universities' official accounts, student associations' accounts, etc
Government	usually referring to local authorities
Hall of Fame	famous individuals from all of walks of life
Media	usually referring to news agencies, television broadcasters, even we-media
Pioneer	grassroots, usually taking an active part in Sina Weibo, whose identities have been successfully verified by Sina Weibo
Website	a window on a website, usually giving an absorbing summary through Sina Weibo but linking details to its own webpages
Weibo Girl	girls who are addicted to sharing selfies with others and who use Sina Weibo as a platform to promote themselves

(Knoke and Yang, 2008). The fundamental principle behind the data gathering is to find friends of friends (FoF). This will be further explained in the following sections.

5.2.1 Social Relations

There are 4 types of social connections on Sina Weibo. Given two users $v_{1,1}$ and $v_{1,2}$ (see Figure 5.2), we have:

- neither $v_{1,1}$ nor $v_{1,2}$ follow each other
- $v_{1,2}$ follows $v_{1,1}$ but $v_{1,1}$ does not follow back
- $v_{1,1}$ follows $v_{1,2}$ but $v_{1,2}$ does not follow back
- $v_{1,1}$ and $v_{1,2}$ follow each other

Type 4 is the only symmetric relation. In this case, $v_{1,1}$ is a friend of $v_{1,2}$ and $v_{1,2}$ is also a friend of $v_{1,1}$. Type 4 is considered a bilateral friendship. In this research work, bilateral friendships were used to collect data from Sina Weibo.

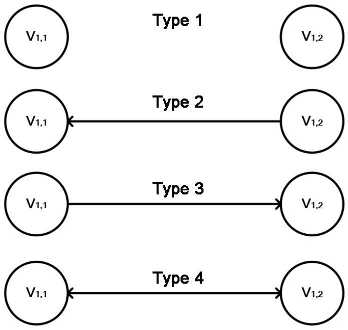
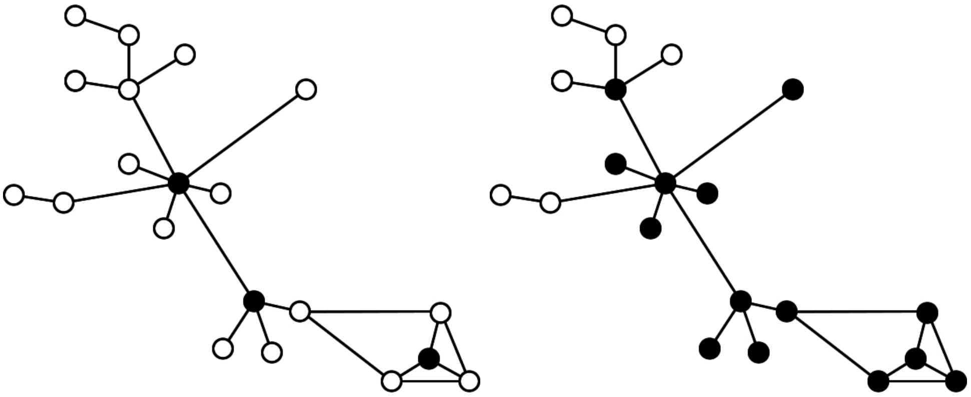


Figure 5.2: Four types of social relations on Sina Weibo

5.2.2 Search Algorithm

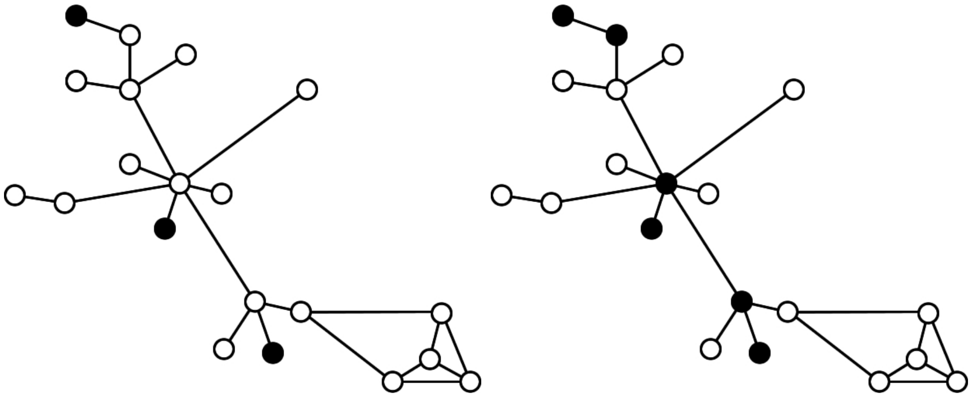
Breath-first search (BFS)(see Section 3.2) is a well-known graph traversal algorithm that has been widely used as a crawling strategy to collect data from social networks (Catanese *et al.*, 2010; Chau *et al.*, 2007; Gjoka *et al.*, 2010; Mislove *et al.*, 2007; Wilson *et al.*, 2009; Ye *et al.*, 2010). A BFS program was written and used to collect data from Sina Weibo. The traversal starts from a set of vertices (i.e. seeds) and continues by visiting the vertices adjacent to the last vertices (see Figure 5.3).



(a) Starting from vertices in black (b) Visiting the adjacent vertices and colouring them black

Figure 5.3: An efficient traversal where 11 vertices were visited after the first iteration

It is worthy mentioning that selecting proper seeds is important. For example, using the seeds with low degree (i.e. having few connections) (see Figure 5.4) as the starting points can lead to an inefficient traversal (e.g., only 6 vertices were visited after the first iteration), compared to the seeds with high degree (e.g., 11 vertices were visited after the first iteration).



(a) Starting from vertices in black (b) Visiting the adjacent vertices and coloring them black

Figure 5.4: An inefficient traversal where 6 vertices were visited after the first iteration

5.2.3 Choice of Seeds

Table 5.2 lists 6 accounts used as the starting points for data collection. As the screen names include Chinese characters, for convenience, their domain names on Sina Weibo were used.

Three 'Pioneer' accounts from the top three cities in China were selected. Recent research shows that people are still bounded by physical distance even though the earliest social networking service appeared decades ago (Backstrom *et al.*, 2010). Thus, picking one 'Pioneer' user from Beijing (Northern China), one 'Pioneer' user from Shanghai (Eastern China) and one 'Pioneer' user from Guangzhou (Southern China) reduced the possibility of finding existing users that have been visited before. The reason the other three accounts 'hejiong', 'panshiyi' and 'people's daily' were selected, is that they have many connections with other verified users from all walks of life, for example, 'hejiong' has more than 500 friends (i.e. bilateral friendships)

Table 5.2: Choices of seeds

Account	Description
hejiong	A famous anchor, who has many connections with other celebrities. He has more than 50 million followers.
panshiyi	A business magnate, who is the chairman of SOHO China, the largest prime office real estate developer in China. He has more than 17 million followers.
rmrb	An offical newspaper of the government of China, a giant in mass media. It has more than 28 million followers.
haroldlee	An ordinary person who works at a consulting company. He lives in Beijing.
wraithree	An ordinary person who works at a shopping center. She lives in Shanghai.
jerjj	An ordinary person who works at an IT company. He lives in Guangzhou.

including 'Hall of Fame' accounts, 'Brand' accounts, 'Campus' accounts, 'Government' accounts, etc. Most of his friends also have many connections with others. Thus, using them as the starting points increased the possibility of finding new users from all walks of life, and eventually created a ripple effect and more efficient traversal.

5.2.4 Crawler for Social Networking Services

'Crawler' is a generic term for any program used to automatically discover and scan websites by following links from one webpage to another (Rosenfeld, 2002). As most SNSs use Dynamic HyperText Markup Language (DHTML) (i.e. a generic term for any technologies used to create web pages that are not static web pages) to make their websites more lively, using a Web crawler to retrieve information from a combination of markup tags and programming scripts is not easy. Also, it involves heavy Web traffic, e.g., in order to get the list of friends of a user, the Web crawler needs to explore several more pages when one page is not enough to list all the friends. Instead, the official API is used to extract data from the server. Sina Weibo provides a REST API for third-party developers. The crawler designed for use in this research work was built on the official API. This made the implementation more efficient. For example, in order to get the list of friends of a user, the crawler only needs to send a request to the server. As shown in Figure 5.5, 6 crawlers were used simultaneously to collect data from Sina Weibo. All the data collected was then integrated into one database.

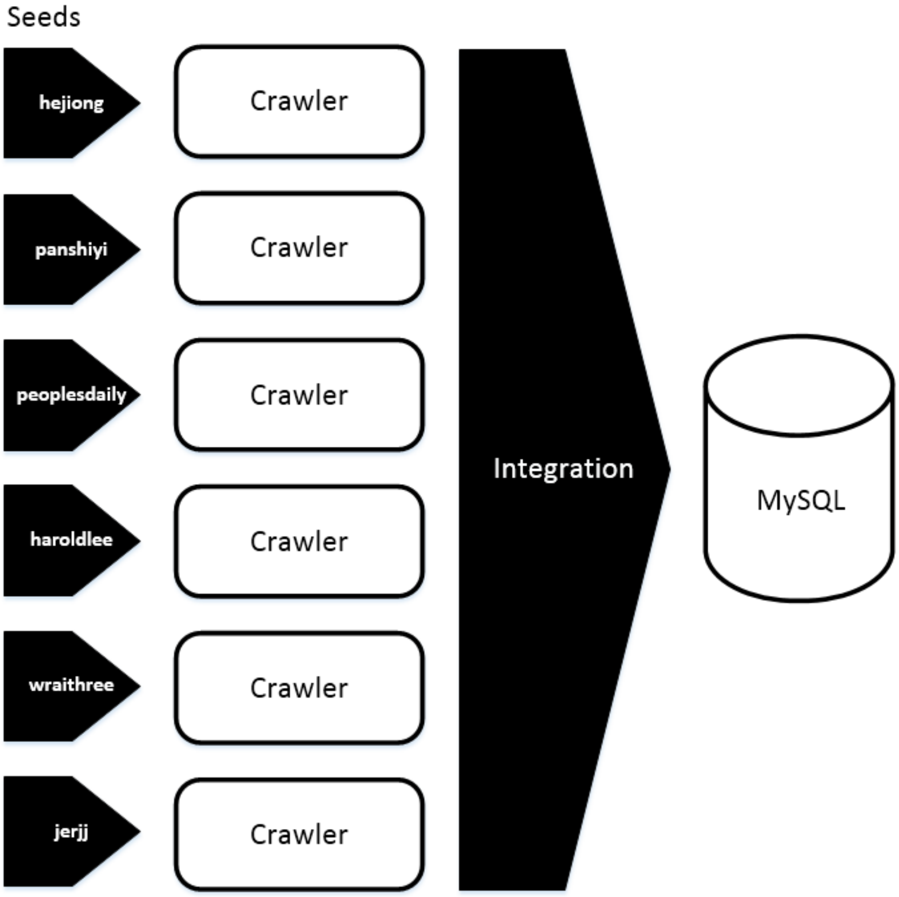


Figure 5.5: Six crawlers were deployed at one time

5.2.5 API Access

The official API is used to acquire the user profiles and the social connections between them. OAuth authentication is required when trying to access to the REST API. An access token (i.e. a permit) is granted to the crawler (see Figure 5.6) once an OAuth request is authorised by the resource owner. The crawler uses the token to access to the resources protected by default.

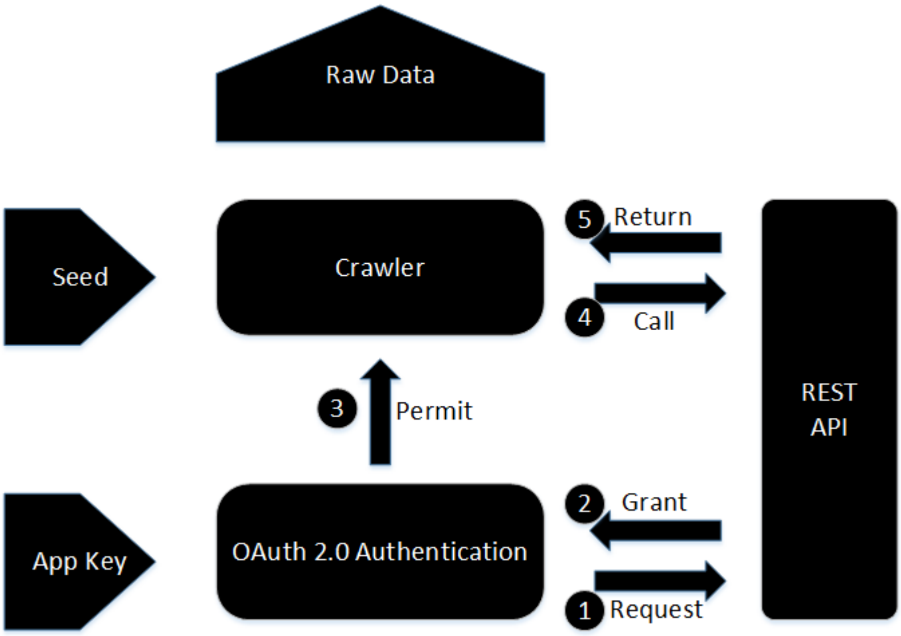


Figure 5.6: The mechanism of how to access to the API

The API only allows the third-party application to make a limited number of calls per hour (see Table 5.3). It is worthy mentioning that a maximum of 150 calls are allowed per hour, regardless of what API function is being called. Because of the rate limits, it is a time consuming task to acquire millions of user profiles and the social connections between them. This is why 6 crawlers were deployed at one time, with each of them having its own App Key (i.e. a string used to identity the application when making requests to the API). In an ideal case, 900 calls can be made per hour.

Table 5.3: An example of rate limits on Sina Weibo

API function	**Number of calls per hour**	**Number of calls per day**
statuses/update	15	50
statuses/repost	15	50
friendships/create	15	50
users/show	150	N/A
statuses/count	150	N/A
friendships/friends/bilateral	150	N/A

5.2.6 Data Structure

Crawlers were used to acquire user profiles and the social connections between them. A user profile was defined as consisting of the following attributes (see Table 5.4). Because little data is in the necessary format, extra computation and calls were required to make the data 'useful'. For example, in order to calculate the number of comments a user has received so far, the crawler had to retrieve the number of comments of each weibo the user posted on Sina Weibo and then add them up. A data set of over 1 million user profiles has been created for this study.

Table 5.4: User profile

Attribute	**Description**	**Example**
uid	a unique number (8 digits) assigned to a user profile	12144623
screen name	the name a user chooses to use for communicating with others online	Jerry Xu Xu Xu
province	the province where a user lives	Hebei
city	the city where a user lives	Guangzhou
gender	male or female	Male
followers	the number of followers a user has	256
followees	the number of followees a user has	320
friends	the number of friends a user has	125
weibo	the number of weibo a user has posted on Sina Weibo	1200
comments	the number of comments a user has received so far	1200
reposts	the number of times a user has been retweeted	16
likes	the number of times a user has been liked	632
verified type	the type of verification a user belongs to	Pioneer
weiage	the number of years a user has used Sina Weibo	4

The data structure illustrated in Table 5.5 was used to store the social connections. The 'uid' on the left hand side (LHS) always follows the 'uid' on the right hand side (RHS). The social connections that were stored are those between users whose profile data had been stored.

Table 5.5: Data structure for social relations

LHS	RHS
1065517411	2651153623
2864652252	2104908771
1889636460	2105665795

An example is shown in Figure 5.7. Users whose profile data had been stored (i.e. $v_{1...16}$) are coloured in black and users whose profile data had not been stored (i.e. v_7, v_8 and v_9) are coloured in white, then the social connections $\langle v_1, v_4\rangle$, $\langle v_4, v_1\rangle$, $\langle v_2, v_4\rangle$, $\langle v_3, v_4\rangle$, $\langle v_5, v_4\rangle$ and $\langle v_6, v_4\rangle$ are retrieved, and $\langle v_6,\ v_7\rangle$, $\langle v_6,\ v_9\rangle$, $\langle v_7,\ v_9\rangle$, $\langle v_9, v_7\rangle$ and $\langle v_9, v_8\rangle$ are ignored.

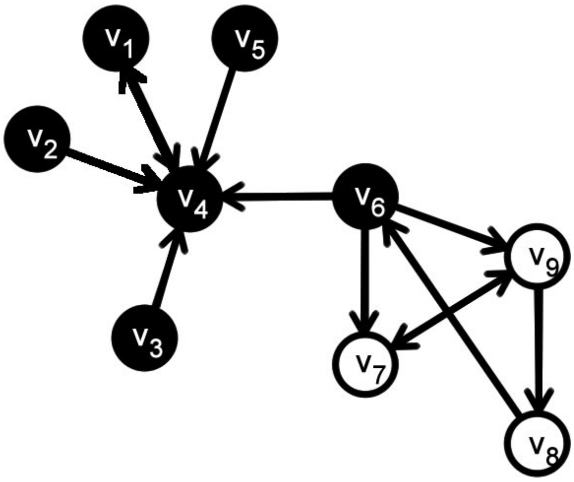


Figure 5.7: Retrieving social connections between the selected users. Users coloured in black indicate that the profile data had been stored. Users coloured in white indicate that the profile data had not been stored.

5.3 Data Sampling

Data sampling is necessary, as otherwise the data of 1 million users would have been too big to be handled within a reasonable time. Considering a graph of 1 million vertices, the number of edges reaches 500 billion for a complete graph (i.e.

every pair of vertices is connected by an edge) and 500 million for a random graph (Erdös and Rényi, 1959) where all pairs of vertices are connected with probability 0.001. Estimating the number of social connections based on the number of users in a social network is more complex than estimating the number of edges based on the number of vertices in a random graph. Although the probability that two users have a social connection between them is hard to measure, an understanding of the volume of the data for the 1 million users that were being used in this research work is still possible. Manipulating a graph of such size is a big challenge considering the computing hardware available. Even though many tasks such as community detection require only linear processing time with respect to the number of edges, they are still time consuming, considering the graph has hundreds of millions of edges. The storage requirements are even more demanding (Fortunato, 2010).

As simple random sampling is not able to reflect the makeup of the population because of the randomness of the selection. Stratified sampling was used in this research work, where samples were drawn based on the distribution of different verified types of users in Sina Weibo. According to the data given by the development team of Sina Weibo, 99.0054% of users (i.e. 729978536 out of 737311501 users) were unverified; 0.1467% of users (i.e. 1081621 out of 737311501 users) were 'VIP' users, including 'Agency', 'Application Software', 'Brand', 'Campus', 'Government', 'Hall of Fame', 'Media', 'Website' and 'Weibo Girl'; and 0.8479% of users (i.e. 6251344 out of 737311501 users) were 'Pioneer'.

5.4 Data Clean-up

Social networks have the properties of being large scale but low density. As shown in Table 5.6, a graph of Facebook with 721.1 million of users only had 68.7 billion edges, where the density of the graph was 2.641×10^{-7} (Backstrom *et al.*, 2010). A graph of Twitter (Kwak *et al.*, 2010) had a density of 8.45×10^{-7}, which is also very low. As has been well-recognised, data sparsity complicates analytic computation because it makes the problem much noisier (Newman and Girvan, 2004).

Table 5.6: Social networks of sparse graph

Social network	**Vertices**	**Edges**	**Density**
Facebook	721.1 million	68.7 billion	0.0000002641
Twitter	41.7 million	1.47 billion	0.000000845

In this research work, two methods were used to improve the density of the data. First, bilateral friendships were used to expand the search of users. It eliminated zombie accounts (i.e. fake or artificial accounts, most time, used for spamming) because zombie accounts are rarely followed back by the other users. Second, a better connected graph such as the largest connected graph was used as the input of the experiments. The largest connected graph provided a graph in which any two users were connected to each other directly or indirectly. Isolated users were removed because they were not connected in any way.

Chapter 6

The Geographies of Communities

This chapter conducts a case study to investigate the relationship between the geographic diversity of communities and the social diversity of communities.

6.1 Data Preparation

Sample 1A is a sample of 100,000 users that were randomly and proportionally chosen from SinaData from 3 different types of users (i.e. 99% from ‘Unverified’, 0.85% from ‘Pioneer’, and 0.15% from ‘VIP’). Stratified sampling ensures every subgroup of the population is sampled according to its relative size.

A directed graph for Sample 1A, D1A, was created. D1A consists of 100,000 vertices and 1,210,469 edges, where vertices represent users and edges represent the social connections between the users. The direction of the arc indicates the relation of who follows whom.

D1A satisfies the properties of a power law distribution, a small-world phenomenon and strong community structure.

Figure 6.1 clearly shows a power law distribution, such that most of the vertices have a relatively small number of in-degree (80% of vertices have in-degree less than or equal to 10) while a small group of vertices have very large in-degree. The vertex that has the maximum in-degree has 22483 arcs incident to it.

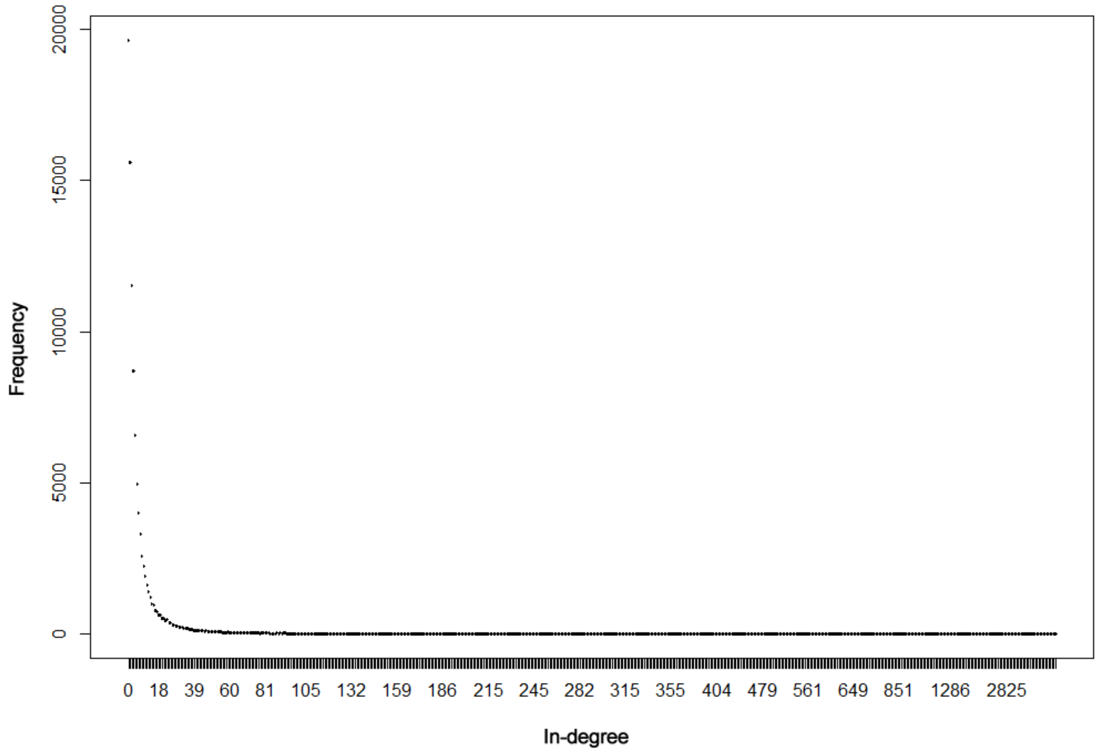


Figure 6.1: The in-degree distribution of the vertices in D1A

In order to mitigate the complexity of the computation and strengthen the connections between users, D1A was transformed to an undirected graph, U1A. Two vertices in U1A are classified as being connected to each other if and only if they are connected in both direction in D1A. In the context of Sina Weibo, that means a bilateral friendship between two users. U1A has 100,000 vertices and 231,496 edges. The number of edges in U1A was reduced to approximately one-fifth of the number of edges in D1A. As a consequence, the time and memory taken for SNA were greatly reduced. As users have tighter connections in a network with symmetric relationships (i.e. an undirected graph) than a network with asymmetric relationships (i.e. a directed graph), using an undirected graph for SNA increased the effectiveness of the analysis.

Table 6.1: Strongly Connected Components (see Definition 3.1.9) in U1A

Component size	**Frequency**
1	26339
2	1239
3	257
4	90
5	26
6	12
7	7
8	5
9	1
10	2
37	1
69695	1

U1A has 26,339 vertices that are not connected in any way. That means, either their social connections were not retrieved, they are isolated from the others or they are connected in one direction only. U1A has 27,980 connected components that are strongly connected (see Table 6.1). Because connected components, by nature, are isolated from each other, vertices in a connected component are irrelevant to the vertices in the other connected components. Thus, in order to ensure the findings of this case study have internal connections with each other and the data is relatively big enough, the largest connected component, LCC1A (see Figure 6.2) was used as the input of this case study. LCC1A consists of 69,695 vertices and 228,671 edges. LCC1A keeps about 69.70% of vertices and 98.79% of edges from U1A, meanwhile, the computational time is reduced significantly. For example, identifying communities in LCC1A by InfoMap (see Section 3.3.2), required about 424.09 seconds in terms of the CPU time charged for execution, compared to D1A, which took about 1643.31 seconds (i.e. about 4 times longer than that in LCC1A).

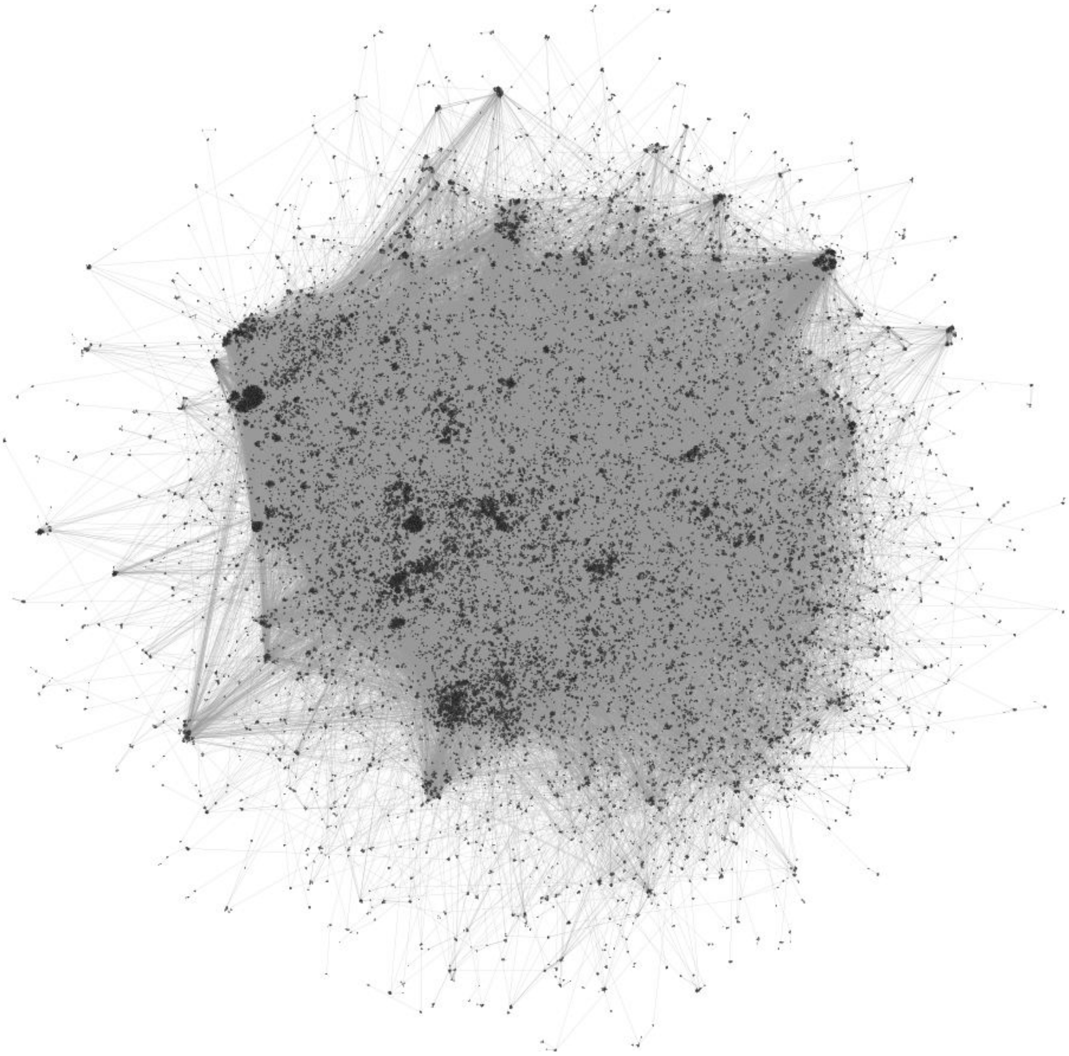


Figure 6.2: LCC1A gives a relatively dense graph compared to U1A. LCC1A also shows strong community structure, where vertices in the dark areas are likely to form new communities.

LCC1A was used as the initial input of SNAHOC. LCC1A was at the bottom level of the hierarchy of communities, in which vertices represent users and edges represent the social connections between them. As shown in Section 4.2.2, if the hierarchy of communities is represented by a rooted tree, vertices in LCC1A are the leaves of the tree. Blondel et al.'s method (see Section 3.3.1) was used in this case study to create the hierarchy of communities for multilevel analysis. Blondel et al.'s method only works with undirected graphs but it is much faster than InfoMap. For example, identifying communities in LCC1A by Blondel et al.'s method, required about 1.02 seconds in terms of the CPU time charged for execution, compared to Infomap, which took 424.09 seconds.

SNAHOC created an abstraction of LCC1A, LCC1A-1, by contracting a group of vertices (i.e. a community) into a single vertex, in which the connections inside the communities (i.e. the connections between the users within the same community) were ignored (see Figure 6.3). This made the connections between the communities (i.e. the connections between the users in different communities) stand out. It is worth noting that there is one and only one edge between each pair of communities if users from different communities are connected to each other. According to the definition given in Chapter 4, LCC1A and LCC1A-1 are two independent graphs, manipulating one graph does not affect the other one (i.e. loosely coupled). The transformation from LCC1A to LCC1A-1 was done by a filter in which information was aggregated based on the purpose of the analysis (see Figure 6.4). Using a rooted tree to represent the hierarchy of communities, each vertex in LCC1A has one and only one parent in LCC1A-1. Vertices in LCC1A that use the same vertex in LCC1A-1 as the parent are the members of the community the parent in LCC1A-1 represents.

6.2 Levels of Abstraction

There are 3 levels of abstraction that were created so as to investigate the influence of the geographic diversity of a community on the social diversity of the community

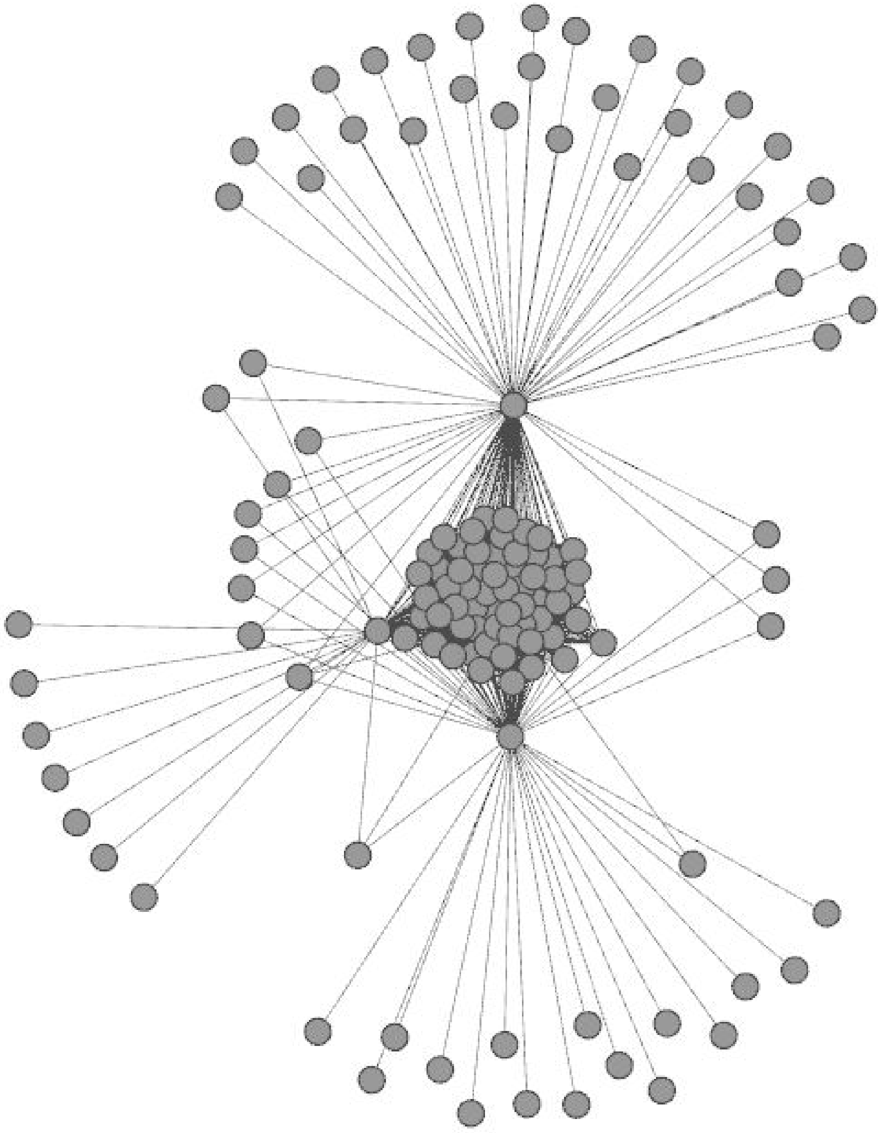


Figure 6.3: LCC1A-1 is an abstraction of LCC1A, where vertices in LCC1A-1 represent communities in LCC1A and edges in LCC1A-1 represent the connections between the communities in LCC1A. The attributes of the vertices describe the characteristics of the communities the vertices represent.

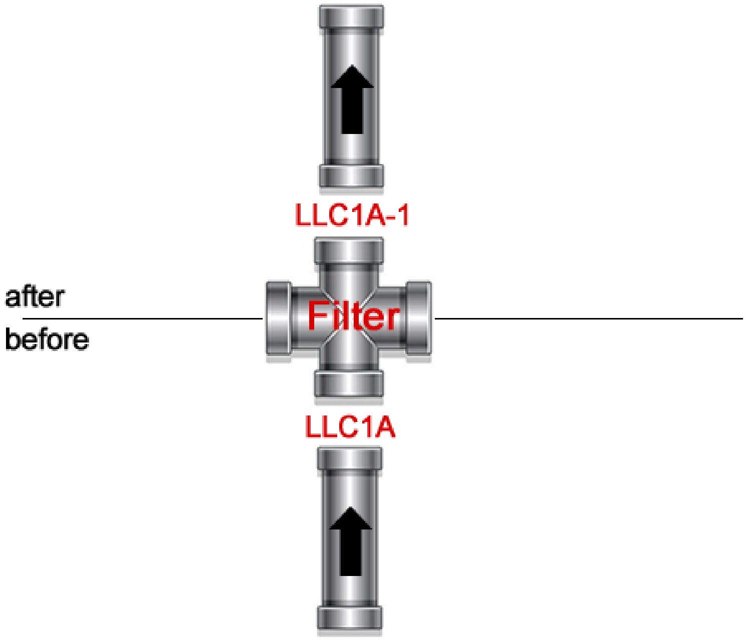


Figure 6.4: LCC1A was piped in and LCC1A-1 was piped out. Information was aggregated based on the purpose of the analysis from bottom to top.

(i.e. the vertex degree)(see Figure 6.5). At the beginning, LCC1A was used as the original input of SNAHOC in which vertices represent individual users. Then, LCC1A-1, LCC1A-1-1, LCC1A-1-2 and LCC1A-1-3 were created as an abstraction of LCC1A, in which vertices represent communities identified in LCC1A. LCC1A-2 was created as different abstractions of LCC1A-1, in which vertices represent communities of communities (i.e. societies). After that, LCC1A-2-2 was created by taking LCC1A-1-2 and LCC1A-2 into account and LCC1A-2-3 was created by taking LCC1A-1-3 and LCC1A-2 into account. Information was aggregated when passing one graph from a lower level of abstraction to a higher level of abstraction.

6.3 The Geographic Diversity of Communities

LCC1A was piped into a filter in which users' location information (i.e. $A_i(province)$ where i refers to the users in Sample 1A) was aggregated and LCC1A-1-1 was piped out where the distribution of locations for each community was displayed by a

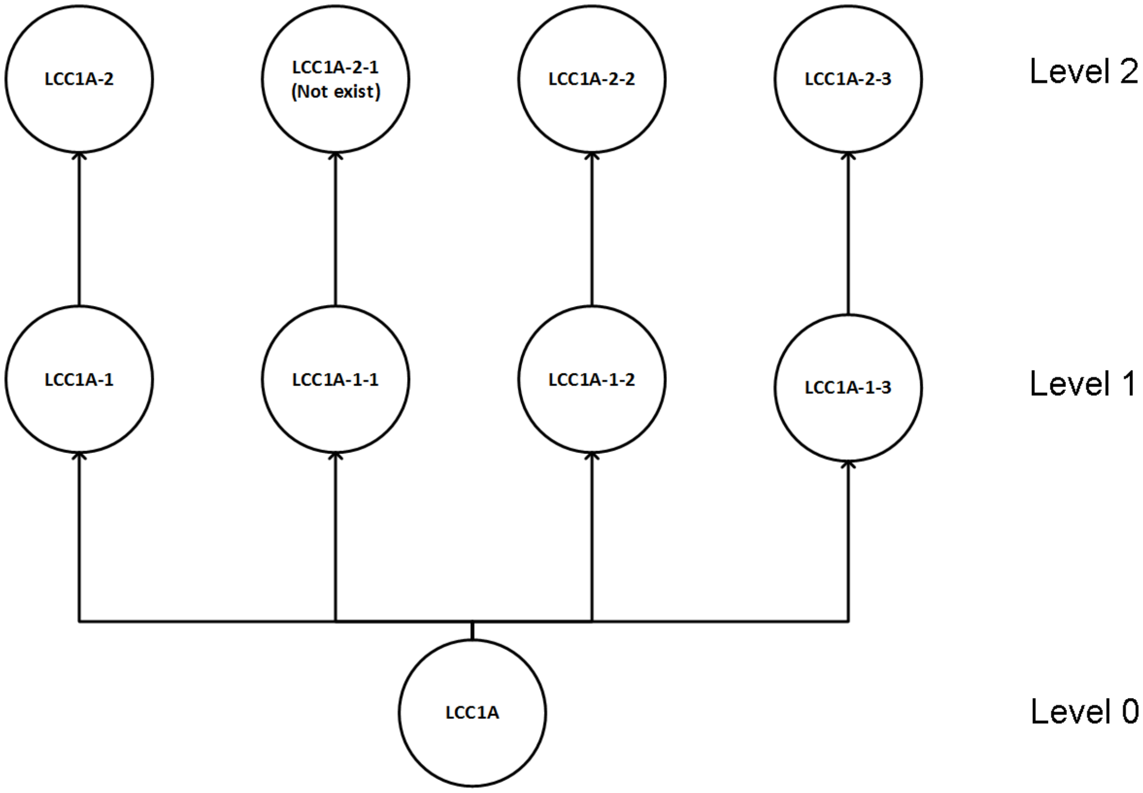


Figure 6.5: SNAHOC created multiple levels of abstraction of LCC1A based on the purpose of the analysis.

pie chart. Compared to LCC1A-1 mentioned above, LCC1A-1-1 is topologically equivalent to LCC1A-1 but presented in a different way (see Figure 6.6).

In LCC1A-1-1, vertices are replaced by pie-charts and labelled by Community IDs. Different colours represent different provinces. Because there are 35 different locations in Sample 1A, when giving each location a different colour it is nearly impossible to avoid using similar colours (e.g., yellow and gold) to represent different locations. Therefore, perceiving the differences between similar colours is not easy, especially, when they are shown in relatively small pie-charts. Thus, the colourfulness of pie-charts was used to investigate if a location dominates a community. Obviously, vertices which are better connected are more colourful (e.g., the ones that are pulled towards the centre of the figure) than vertices which are less well-connected (e.g., the ones that are scattered over the edge of the figure). As shown in Figure 6.6, in most cases, vertices are dominated by a single colour (i.e. taking more than half of the pie), no matter how well they are connected (e.g., vertex 24, 89 and 112) or remote they are from the centre (e.g., vertex 38 and 56). The results presented here also corroborate previous studies (Scellato *et al.*, 2010; Batty *et al.*, 2012) that the members of a community are geographically close to each other.

LCC1A then was piped into a filter in which users' location information was aggregated and LCC1A-1-2 was piped out where the richness of locations for each community was displayed (see Figure 6.7). Both LCC1A-1-1 and LCC1A-1-2 (mentioned above) are abstractions of LCC1A. Refer to SNAHOC, both LCC1A-1-1 and LCC1A-1-2 are at the same level of abstraction but were generated by different filters (see Figure 6.5)

In LCC1A-1-2, vertices are labelled by their richness of locations. The richness of locations is the total number of distinct locations of a community. As shown in LCC1A-1-2, in most cases, vertices which are better connected are labelled by bigger numbers from 20 to 35 and vertices which are less well-connected are labelled by smaller numbers from 1 to 5. There is a clear gap between vertices which are better connected and vertices which are less well-connected as shown in Figure 6.7. The richness of locations of the vertices is positively related to the vertex degree with correlation coefficient of 0.98.

LCC1A was also piped into a filter in which users' location information was aggregated and LCC1A-1-3 was piped out where the geographic diversity for each community was displayed (see Figure 6.8). The geographic diversity was calculated by the Shannon-Wiener index (see Section 3.5). The geographic diversity takes not only the richness of locations but also the evenness of locations into account. In LCC1A-1-3, vertices are sized in proportion to their geographic diversity. Even though, in most cases, the less connected vertices are relatively smaller than the better connected vertices, in terms of the size of the vertices (i.e. the geographic diversity of the vertices), there are a few exceptions that have been circled in Figure 6.8. The geographic diversity is positively related to the vertex degree with correlation coefficient of 0.79.

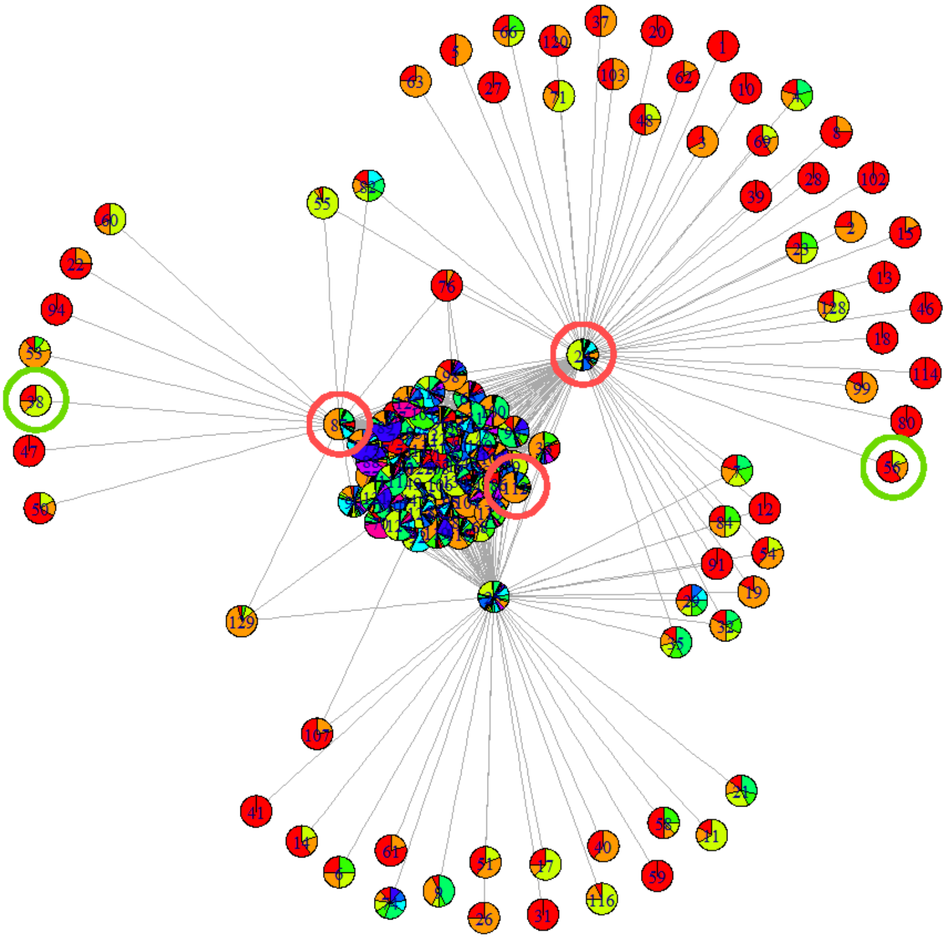


Figure 6.6: LCC1A-1-1 shows if a location dominates a community. In most case, vertices are dominated by a single colour. For example, vertex 24, 89 and 112, circled in red, which have many connections with the others, are dominated by yellow, blue, and orange respectively. Vertex 38 and 56, circled in gree, which have only 1 connection with the others, are dominated by yellow and red.

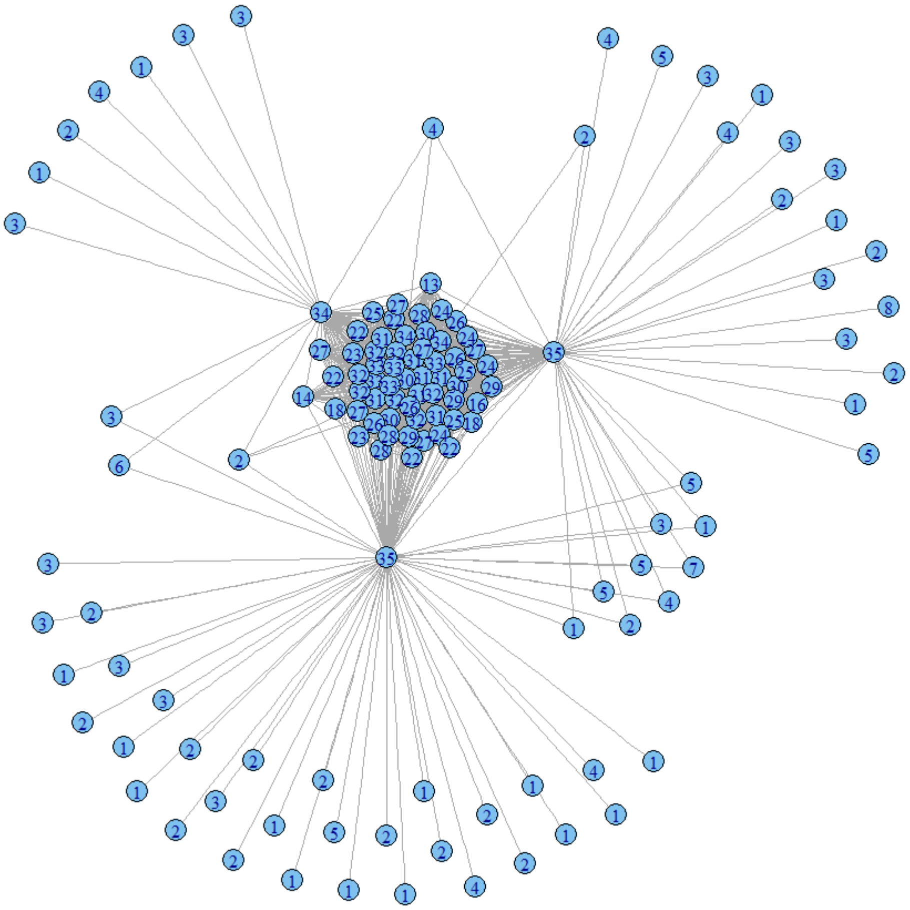


Figure 6.7: LCC1A-1-2 shows a clear gap between well-connected vertices and less well-connected vertices in terms of their richness of locations.

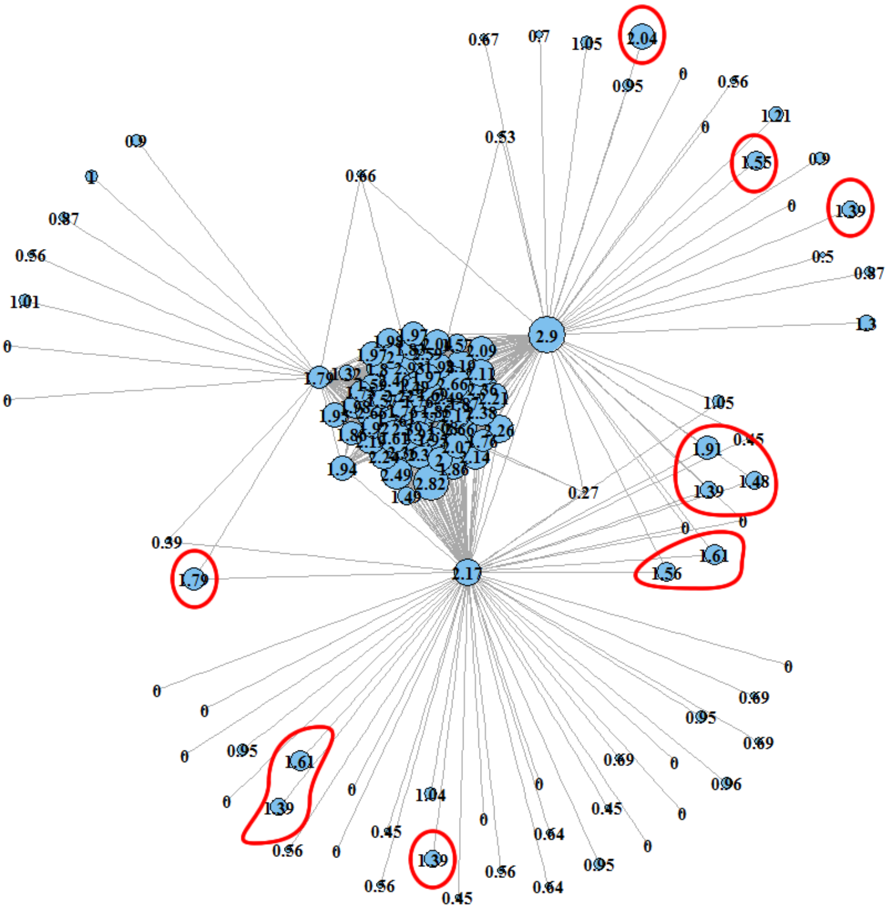


Figure 6.8: Vertices that are less well-connected but have relatively high geographic diversity are circled. The gap between well-connected vertices and less well-connected vertices in terms of their geographic diversity is not that clear compared to the richness of their locations as shown in Figure 6.7.

LCC1A-1, LCC1A-1-1, LCC1A-1-2 and LCC1A-1-3 are all at the same level of abstraction (i.e. an abstraction of LCC1A) but are presented in different ways.

6.4 The Characteristics of Communities

SNAHOC created an abstraction of LCC1A-1, LCC1A-2 by contracting a group of communities into a single society, as shown in Figure 6.9. In LCC1A-2, each vertex represents a society (i.e. a community of communities). Vertices are labelled by Society IDs. The size of the vertices varies depending on the size of the societies. If a member of Society 1 (i.e. a vertex in LCC1A-1 that is clustered into Society 1) has a connection with a member of Society 2, then there is an inter-connection between Society 1 and 2. Multiple inter-connections between the societies are contracted into a single edge in LCC1A-2. Edges are labelled by the number of inter-connections they represent. Compared to inter-connections, the connections within the societies are considered intra-connections. This case study explored the characteristics of the communities identified in LCC1A by grouping them into different societies and observing the inter-connections between the societies and the intra-connections within the societies.

LCC1A-2-2 is an equivalent of LCC1A-1-2 but was created by taking LCC1A-2 into account (see Figure 6.10). Vertices within the same society are bounded by a coloured convex hull (i.e. the minimum convex geometry that encloses all geometries within the set). Different colours represent different societies, Society 1 is red, Society 2 is green and Society 3 is blue. The size of the vertices varies depending on their richness of locations. Vertices of low richness are relatively smaller than vertices of high richness in terms of the size of the vertices. Edges coloured red represent the inter-connections between the societies and Edges coloured black represent the intra-connections within the societies.

As shown in Figure 6.10, vertices of high richness are more likely to connect with others that are in different societies. Vertices of high richness are pulled towards the centre of the figure because they are densely connected to each other compared to

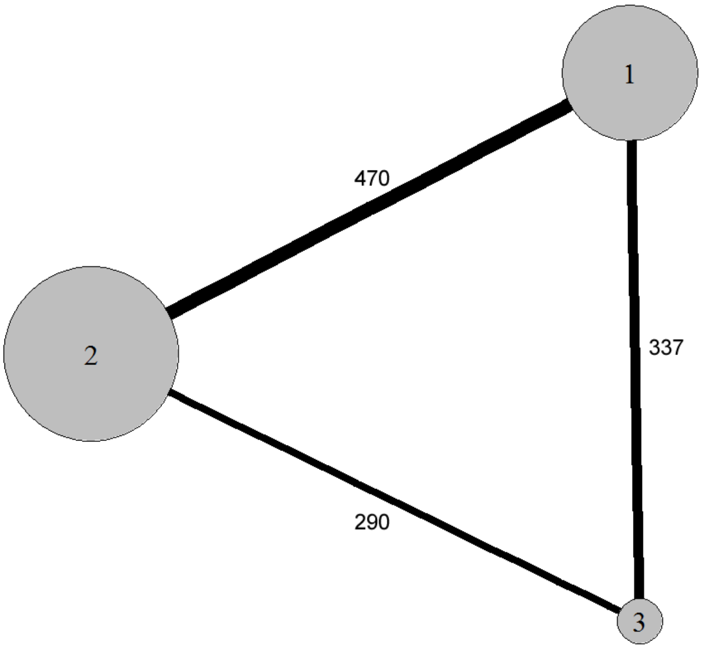


Figure 6.9: 131 communities identified in LCC1A were grouped into 3 societies in LCC1A-2.

vertices of low richness that are scattered over the edge of the figure because most of them are directly connected to a common vertex (circled in Figure 6.10) and every vertex is indirectly connected to every other through the central vertex.

Vertices of Society 3 (i.e. coloured blue) have a richness of 26.0588 on average, which is much higher than vertices of Society 1 (i.e. 16.04 on average) and Society 2 (i.e. 10.5469 on average). Figure 6.11 shows the box plots of the richness of locations for each society. There is more variation in Societies 1 and 2 that both range from 1 to 35, whereas the richness of locations of Society 3 ranges from 18 to 33. 50% of the vertices in Society 1 have the richness of locations less than 19 and 50% of the vertices in Society 2 have the richness of locations less than 3, compared to Society 3, where 50% of the vertices have the richness of locations between 22 and 31. Figure 6.12 shows the frequency histogram of the richness of locations for each society as a supplement to discover the distribution of the richness of locations for each society. Richness in Societies 1, 2 and 3 follows a bimodal distribution in which there are two peaks in the distribution. Even though vertices of high richness (i.e.

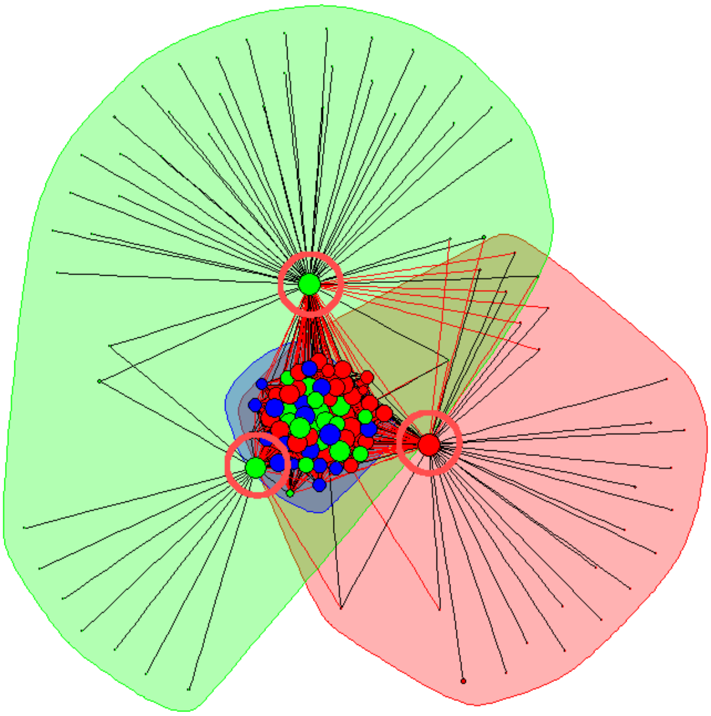


Figure 6.10: Inter-connections in most cases happen between big vertices.

greater than 6, which is the median of all of the richness of locations in Societies 1, 2 and 3 taken together) in Societies 1 and 2 are more than that in Society 3 because of the large base of vertices. Vertices of high richness are much more frequent in Society 3 compared with Societies 1 and 2. Reflecting on Figure 6.10, vertices in Society 3 are much the same in terms of the size of the vertices but vertices in Societies 1 and 2 are either too small or too big (i.e. either extremely low richness or high richness) because of a bimodal distribution (see Figure 6.12).

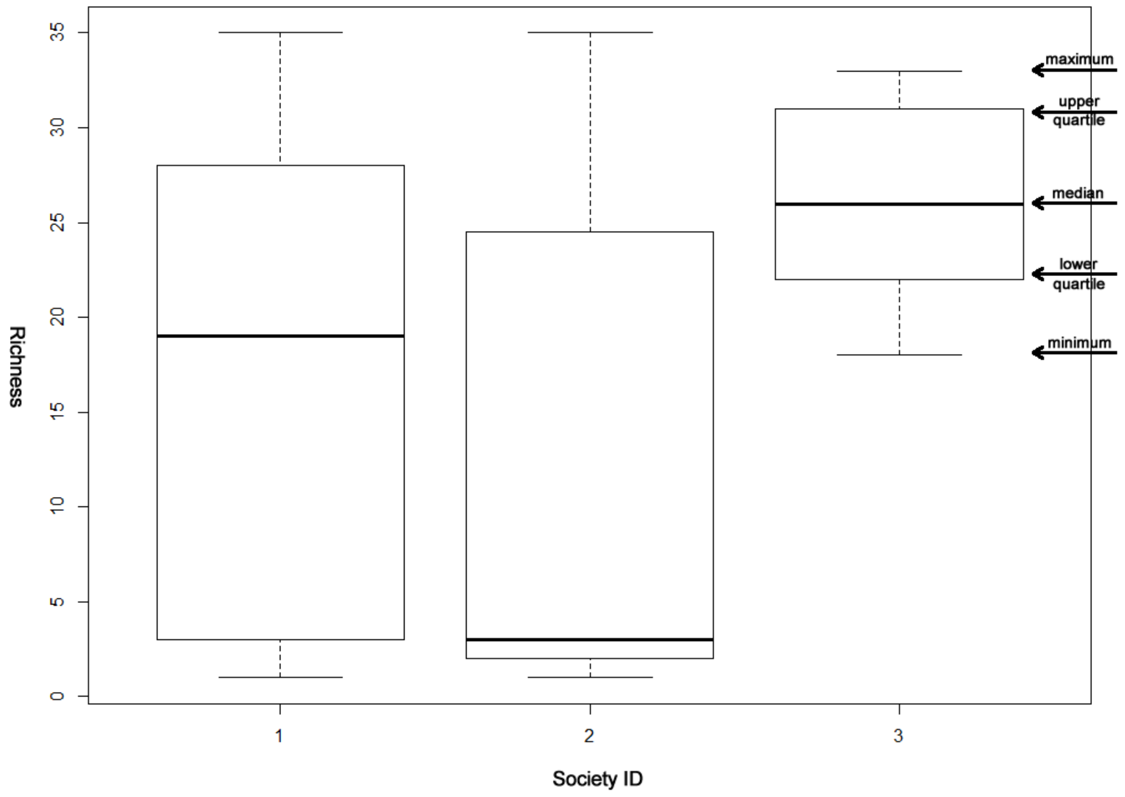


Figure 6.11: The median is the number that 50% of data is greater than it. The upper quartile is the number that 25% of data greater than it. The lower quartile is the number that 25% of data less than it. The maximum is the largest number and the minimum is the smallest number.

Most of inter-connections happen between vertices which have high richness. Reflecting on Figure 6.10, big vertices are more likely to have inter-connections, but tiny vertices in most cases are connected by intra-connections. The correlation coefficients between the richness of locations and the number of inter-connections,

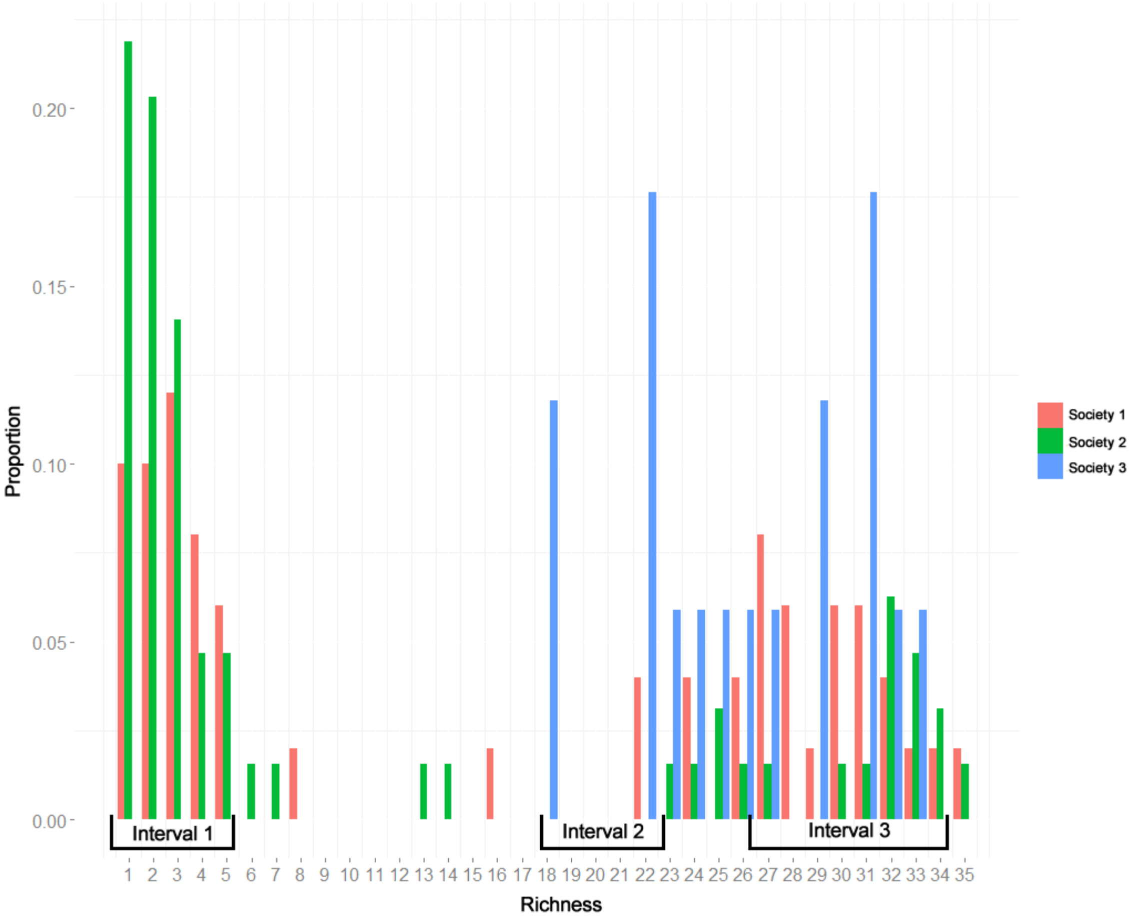


Figure 6.12: The first peak in Societies 1 and 2 appears at Interval 1 and the second peak appears at Interval 3. The first peak in Society 3 appears at Interval 2 and the second peak appears at Interval 3. Vertices of high richness in Society 3 are more frequent than that in Societies 1 and 2.

for each society, are given in Table 6.2, so was the correlation coefficients between the richness of locations and the number of intra-connections.

Table 6.2: The correlation coefficients between the richness of locations and the number of inter-connections (or intra-connections)

Society ID	**Correlation between richness and inter-connections**	**Correlation between richness and intra-connections**
1	0.9756	0.9477
2	0.9902	0.8976
3	0.8189	0.5752

A strong relationship exists between the number of inter-connections and the richness of locations. The correlation between the richness of locations and the number of inter-connections actually reaches 0.8189 in Society 3. This is despite the fact that the richness of locations and the number of intra-connections are relatively weakly related. The richness of locations in Societies 1 and 2 is strongly positive related to both the number of inter-connections and intra-connections; on the other hand, the richness of locations in Society 3 is strongly positive related to the number of inter-connections and weakly related to the number of intra-connections. The richness of locations has greater influence on the number of inter-connections than that of intra-connections in Society 3. Because vertices in Society 3 are more likely to have high richness, compared to vertices in Societies 1 and 2, vertices in Society 3 are more likely to have inter-connections with vertices from other societies. This partially explains why vertices of blue twist together with vertices of red and green in the centre of Figure 6.10.

LCC1A-2-3 is an equivalent of LCC1A-1-3 but was created by taking LCC1A-2 into account (see Figure 6.13). Vertices within the same society are bounded by a coloured convex hull. Different colours represent different societies. In LCC1A-2-2, vertices are sized in proportion to their geographic diversity. Vertices of low diversity are relatively smaller than those of high diversity. Edges coloured red represent inter-connections and edges coloured black represent intra-connections.

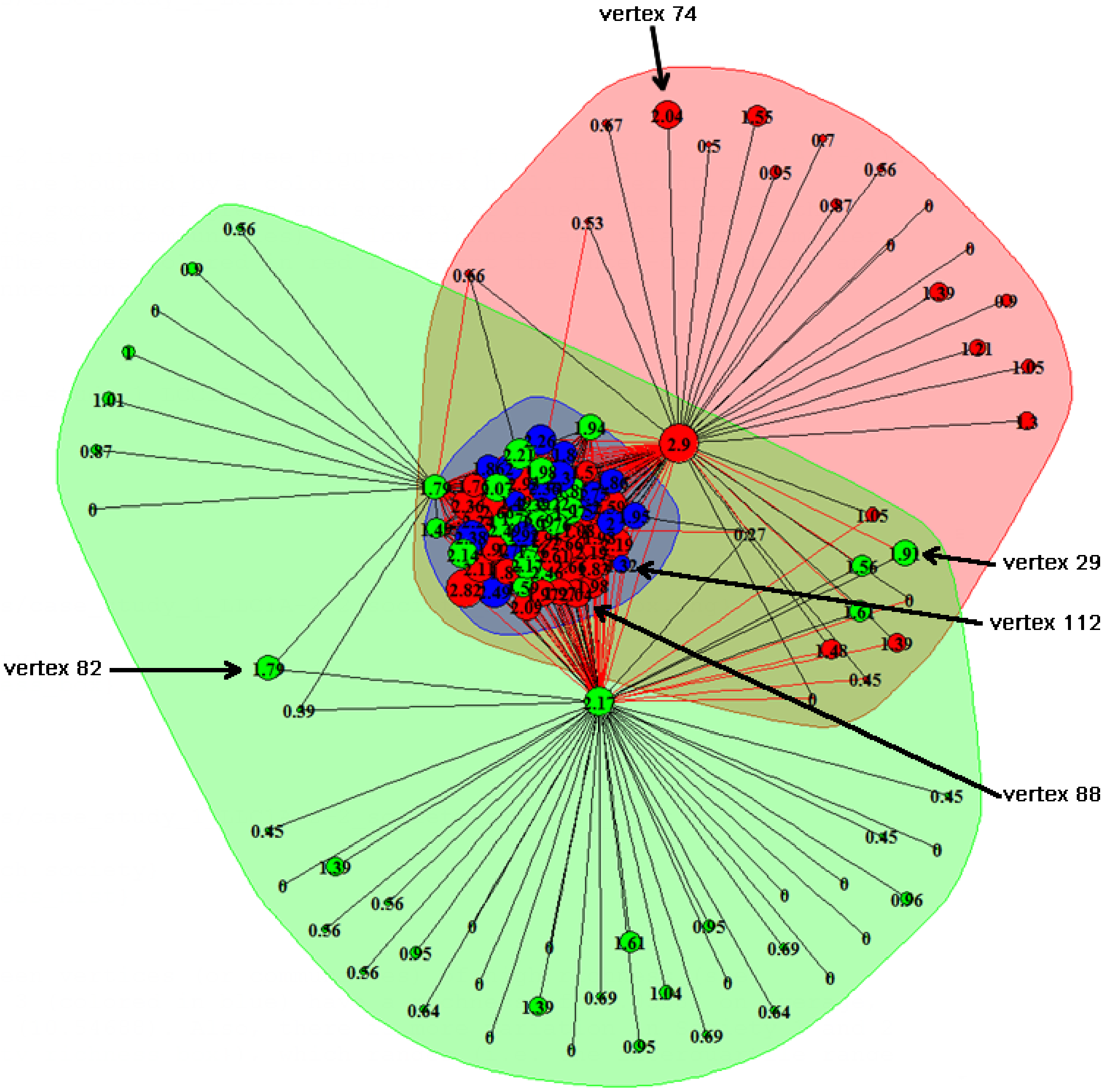


Figure 6.13: Some vertices of high diversity do not have any inter-connections.

Vertices which are less well-connected are much smaller than vertices which are better connected in terms of the size of the vertices in LCC1A-2-2 (see Figure 6.10). However, in LCC1A-2-3, vertices which are less well-connected can still have relatively high diversity (see Figure 6.13). For example, both the vertex 74 and 88 have the same diversity of 2.04 but the former has a degree of 1 and the latter has a degree of 45. Another example is that the vertex 29 has the diversity of 1.91 which is larger than that of 1.32 of the vertex 112 but the degree of the vertex 29 is only 2 which is much smaller than the degree of the vertex 112 which is 43.

The evenness of locations refers to how close in numbers each location is. It is calculated by Equation 3.6. As the calculation of the geographic diversity takes both richness and evenness into accounts (see Equation 3.5), vertices which have low richness but high evenness are still able to gain high diversity. As shown in Figure 6.13, some vertices, such as the vertex 74 and 82, which have the diversity of 2.04 and 1.79 respectively, are connected by intra-connections only. The deduction is that the correlation between the geographic diversity and the number of inter-connections is not that strong compared to that with respect to the richness of locations.

Table 6.3: The correlation coefficients between the inter-connections (or intra-connections) and the SWI for each society

Society ID	**Correlation between diversity and inter-connections**	**Correlation between diversity and inra-connections**
1	0.7990	0.7906
2	0.7755	0.7012
3	-0.3906	-0.2088

The correlation coefficients between the geographic diversity and the number of inter-connections for each society are given in Table 6.3, so are the correlation coefficients between the geographic diversity and the number of intra-connections. It can be seen by comparison with Table 6.2 that the correlation between the diversity and the number of inter-connections is much weaker than that between the richness and the number of inter-connections. Especially for Society 3, there is a weak relationship between the geographic diversity and the number of inter-connections (or intra-connections), in which the number of inter-connections (or intra-connections) decreases as the geographic diversity decreases.

A gradient colour scheme is used in Figure 6.14 to assist with the visualisation of the relative positions of vertices across from blue at the lower end of the number of inter-connections, through purple, to red at the upper end of the number of inter-connections. Figure 6.14 clearly shows that the number of inter-connections is related to the richness only. The colour of the vertices turns to red (i.e. the number of inter-connections increases) moving from the bottom to the top (i.e. from the lower end of the richness to the upper end of the richness). On the other hand, no matter how the evenness is changed, the number of inter-connections does not increase, or decrease accordingly. The colour of the vertices remains blue moving from vertices near the front to those near the back (i.e. from the lower end of the evenness to the upper end of the evenness).

To sum up, the geographic diversity of a community is positively correlated to the number of inter-connections of the community. It is consistent with previous studies (Jehn *et al.*, 1999; Reagans and Zuckerman, 2001) in organisational behaviour that diverse backgrounds do have a positive effect on the social capital of a group (i.e. more contacts outside the group).

6.5 Potential Applications for Commercial Use

Based on the above case study, a potential application is discussed in this section. Assume that the employees of an organisation are on LinkedIn, a social network used for professional networking. Each employee is considered as a vertex and the organisation they belong to forms a community on LinkedIn. SNAHOC can be used to optimise the make-up of the employees of the organisation by analysing the social diversity of the organisation at multiple levels: the micro level (focusing on the geographic diversity within the organisation) and the macro level (focusing on the

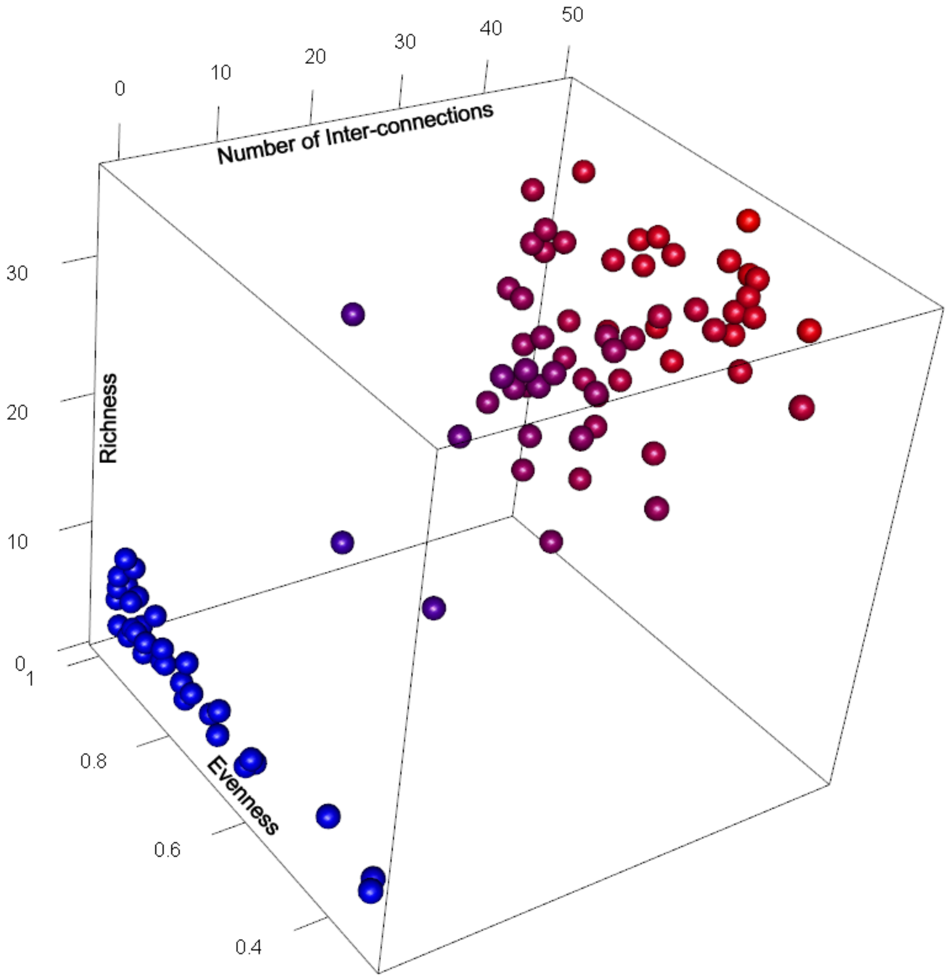


Figure 6.14: The relationship among the inter-connections, richness and evenness.

inter-connections with other organisations). Such information can help the organisation's recruiters to choose the candidates who can increase the geographic diversity of the organisation (so as to foster creative thinking within the organisation) and who can increase the inter-connections with other organisations (so as to expand the organisation's external connections such as marketing channels, etc.).

Chapter 7

The Influence of Communities

This chapter introduces a measure of the influence of a community based on SNA-HOC. The relationship between individual users and communities in terms of influence are examined.

7.1 Data Preparation

Unlike the previous case, in which, an undirected graph was used, the direction of the connection has significant impact on the influence between a pair of users. Thus, a directed graph was used in this case to measure the influence of the users. However, Blondel et al.'s method used in the previous case is not for directed graphs. Instead, InfoMap (see Section 3.3.2) was used in this case to create the hierarchy of communities for multilevel analysis.

Exactly the same sampling strategy was used in this case as in previous one. However, because of the intensive computation encountered in the calculation of PageRank and the high memory usage in directed graphs, it was decided to shrink the sample size from 100,000 users (used in the previous case) to 10,000 users so that the experiments could be done within a reasonable time scale.

Sample 2A consists of 10,000 users randomly chosen from SinaData (i.e. 1,192,972 users), in proportion to the number of users from different types (i.e. 99% from 'Unverified', 0.85% from 'Pioneer', and 0.15% from 'VIP'). There are 15 'VIP' users, 85 'Pioneer' users and 9900 'Unverified' users.

A directed graph for Sample 2A, D2A was generated (see Figure 7.1). D2A consists of 10,000 vertices and 9,814 edges, where vertices represent users and edges represent the social connections between the users. The direction of the arc indicates the relation of who follows whom.

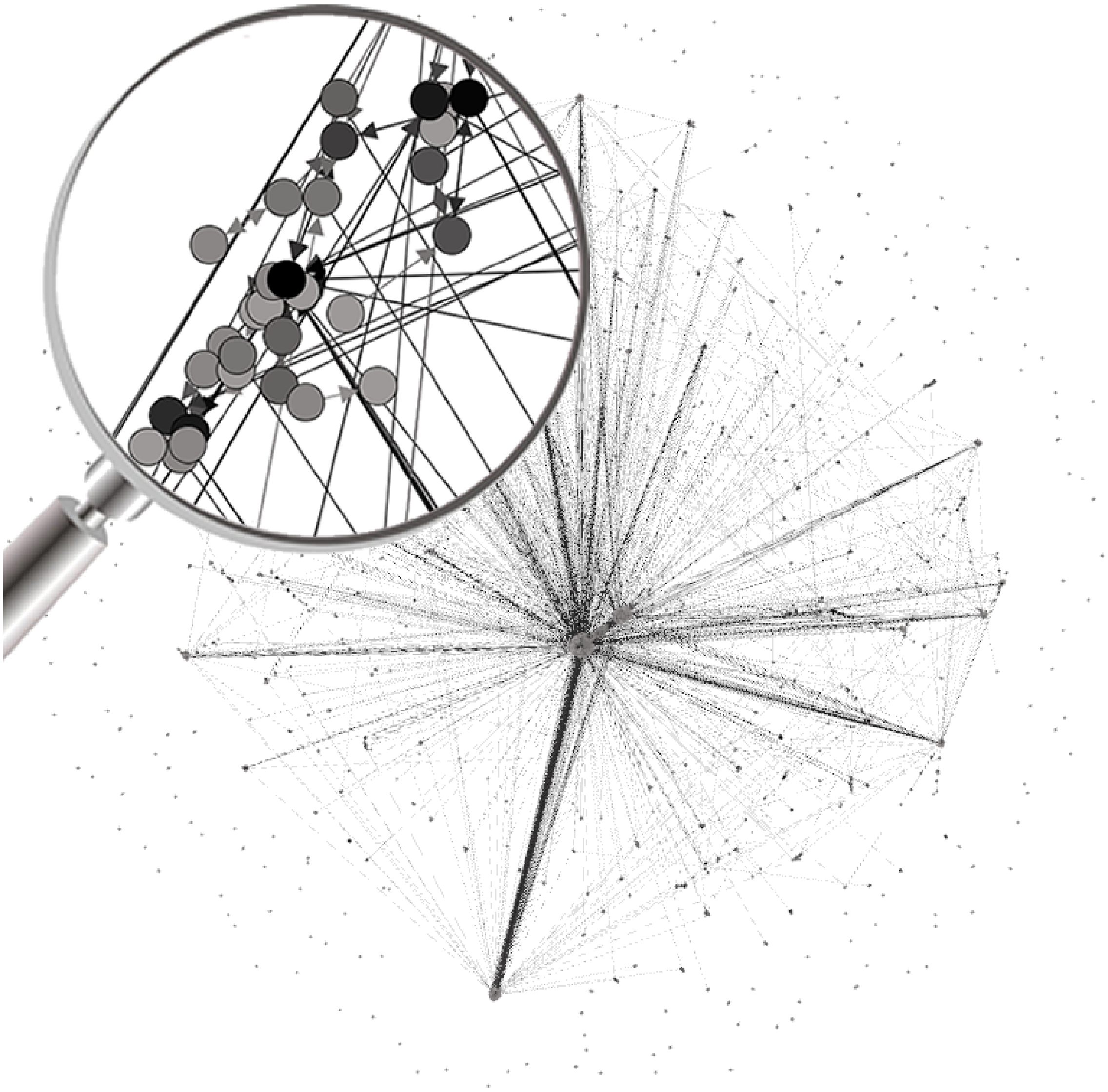


Figure 7.1: Vertices that are pulled together are likely to form communities. In order to better show the directions of the edges in D2A, a part of it was zoomed in.

D2A satisfies the properties of a power law distribution, the small-world effect and strong community structure.

Figure 7.2 shows a power law distribution. 7,686 out of the 10,000 vertices do not have any arc incident to them (i.e. they have an in-degree of 0). 1,562 out of the 10,000 vertices have only 1 arc incident to them (i.e. they have an in-degree of 1). The number of vertices goes down quickly when the in-degree goes up (e.g., 433 vertices have an in-degree of 2, 143 vertices have an in-degree of 3, etc).

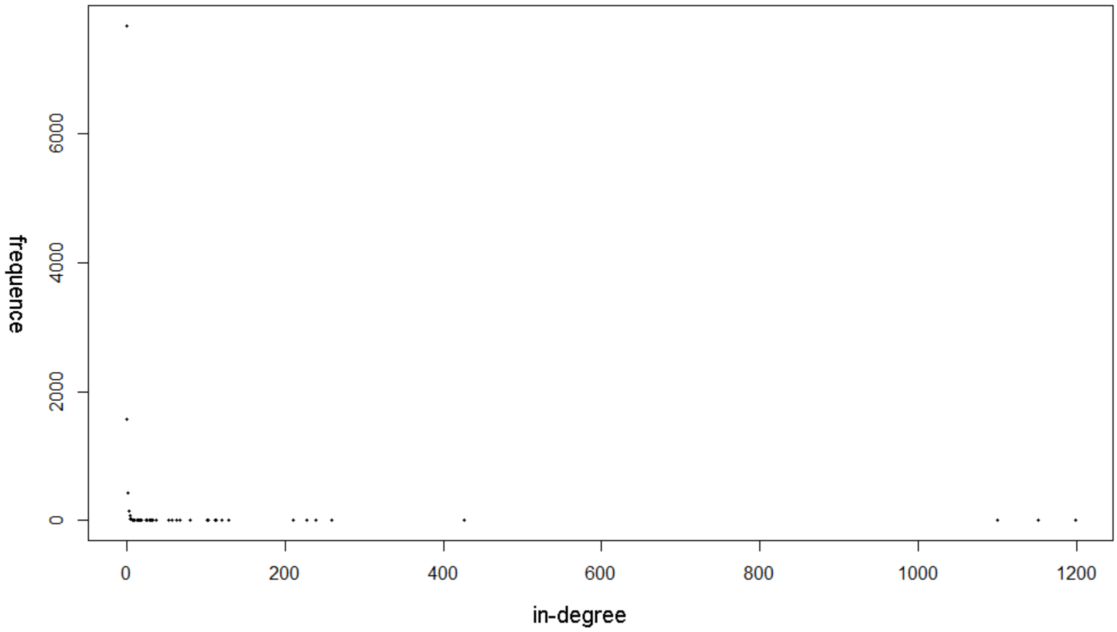


Figure 7.2: The in-degree distribution of the vertices in D2A

The average path length is 6.576. That means users in D2A can be connected to any other users through 6.576 users on average.

D2A also has strong community structure. Vertices that are pulled together (see Figure 7.1) are likely to form communities. Connections within communities are dense but connections between communities are sparse.

D2A has 4,371 vertices that are not connected in any way. Those vertices can be ignored as they can not have any influence on the others and other vertices can not reach them in any way. Thus, RD2A was created by removing them from D2A. RD2A is a directed graph with 5,629 vertices and 9,814 edges.

Although there is no isolated vertices, RD2A still has 211 connected components that are weakly connected. As mentioned in the previous case, connected components are by definition (see Definition 3.1.9) isolated from each other. That means a user's influence in a community can not be passed on to other users in other communities. Thus, the largest connected component, LCC2A was used as the input of this case study. LCC2A consists of 5,157 vertices and 9,382 edges (see Figure 7.3). LCC2A keeps about 91.61% of vertices and 95.60% of edges in RD2A. LCC2A gives a connected network, in which every user is connected to all other users directly or indirectly.

7.2 Influential Users within Communities

In order to calculate a user's PageRank, a weight is added to each edge according to the influence of the initiating vertex. The weight of the edge, $\langle v_i, v_j \rangle$, is calculated as below:

$$W(v_i, v_j) = A_{v_i}(social\_diversity) \times (A_{v_i}(reposts) + A_{v_i}(comments)) \tag{7.1}$$

where $W(v_i, v_j)$ represents the weight of $\langle v_i, v_j \rangle$, $A_{v_i}(social\_diversity)$ is the number of inter-connections v_i initiates, $A_{v_i}(reposts)$ is the number of reposts v_i has been reposted and $A_{v_i}(comments)$ is the number of comments v_i has received. In order to calculate $A_{v_i}(social\_diversity)$, LCC2A was abstracted to LCC2A-1 (see Figure 7.4), in which a group of vertices was contracted into a single vertex (i.e. a community). Instead of Blondel et al.'s method used in the previous case, InfoMap was used. Then, the social connections v_i initiated to vertices that are not in the same community with v_i were counted as $A_{v_i}(social\_diversity)$. A users' PageRank is calculated as below:

$$R(v_i) = \sum_{v_j \in B(v_i)} \frac{R(v_j)}{N(v_j)} W(v_i, v_j) \tag{7.2}$$

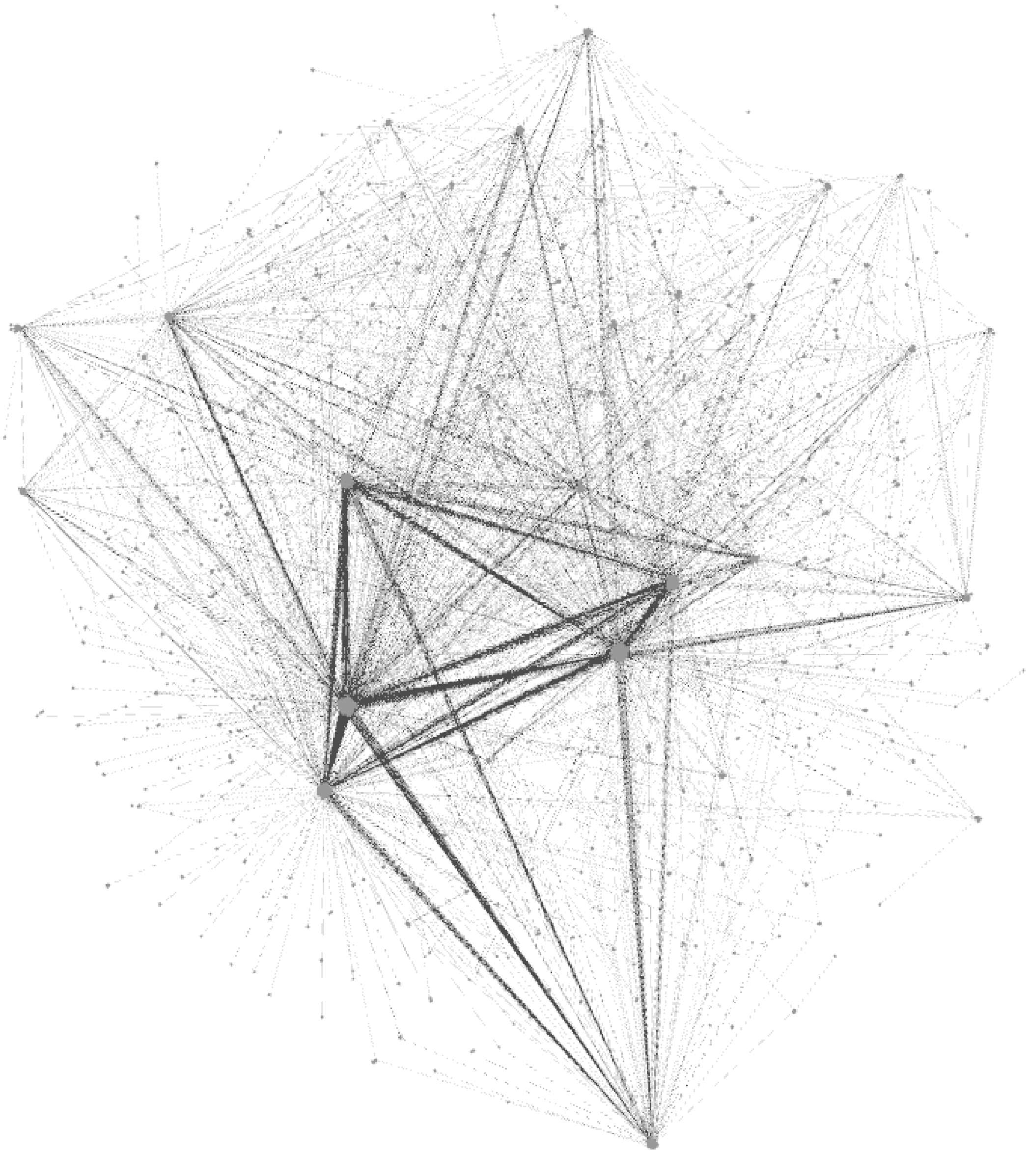


Figure 7.3: Vertices in LCC2A are weakly connected. Compared to D2A, isolated vertices and components were removed. As a consequence, about half of the vertices were removed but most of the edges were kept.

which is a variant of Equation 3.4.

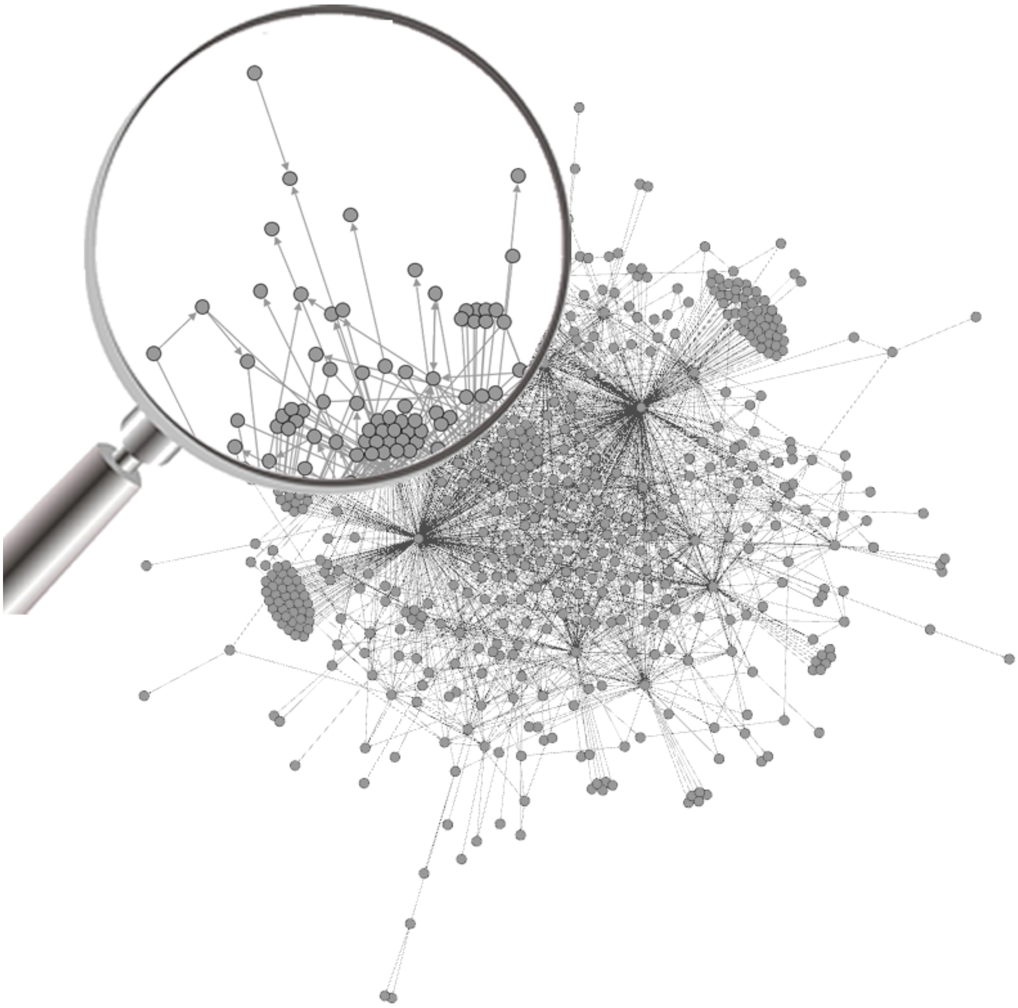


Figure 7.4: LCC2A-1 is a directed graph with 653 vertices and 1,809 edges. Each vertex in LCC2A-1 represents a community.

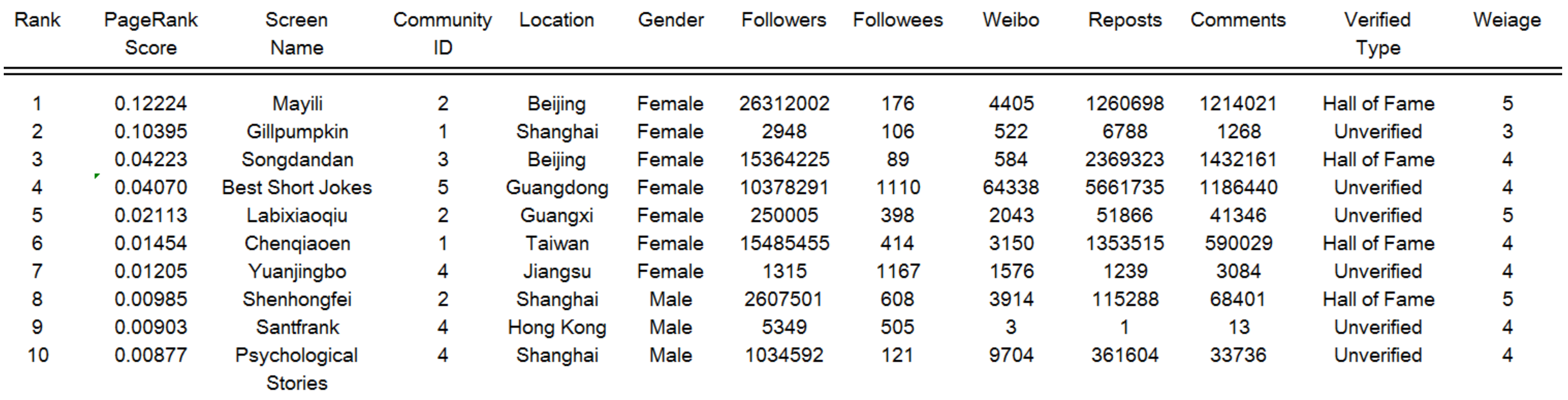
The top 10 influencers are listed in Figure 7.5. 7 out of the top 10 users are female and 3 of those are actresses. 'Best Short Jokes' and 'Psychological Stories' are We-Media. They both have audiences in the millions even though they are unverified accounts. 'Shenhongfei' is a freelance writer and has over 2 million followers. It makes sense that celebrities and We-Media have much more influence than others. However, users like 'Gillpumpkin', 'Labixiaoqiu', 'Yuanjiangbo' and 'Santfrank' are also very influential although they do not have remarkable characteristics. They have high PageRank scores because of the nature of the structure of LCC2A, e.g., 'Santfrank' is influential because 'Labixiaoqiu' is a follower of him. 'Liabixiaoqiu' herself is a hub who is connected with another 48 communities (i.e. extremely diverse). Her weibo was re-posted 5,1866 times and she received 4,1346 comments in total (i.e. highly active), which causes 'Liabixiaoqiu' to be an influencer. The influence of 'Santfrank' is almost totally inheirted from 'Liabixiaoqiu'.

It is obvious that the results of the influential users are affected by the sample that is used. The calculation of a user's influence depends on the active level and the social diversity. In this case study, a user's active level is determined by $A(reposts)$ and $A(comments)$. As those two attributes are heavily depend on other users' reactions, they better reflect to what extent a user participates in Sina Weibo, compared to other attributes such as $A(followees)$ and $A(weibo)$, which are determined by users themselves. The social diversity reflects how wide a user's social circle is. It is determined by the topological structure of the network that is used. It is not necessarily the case that a user who has diverse social connections in a network also has diverse connections in another network.

7.3 Weighted Average Influence of Communities

In order to measure the influence (i.e. PageRank) of a community, individual users were grouped into 4 categories based on their PageRank scores, and the influence of the community was determined by the individuals' PageRank scores, in which weighted means were calculated. Another iteration of the PageRank algorithm was then run. The details are given below.

First, users were grouped using the k-medoids clustering (Kaufman and Rousseeuw, 2009), based on their PageRank scores. In order to determine the optimal number of clusters, 14 possible solutions were evaluated and the optimal solution was chosen by comparing the Sum of Squared Error of Prediction (SSE) each solution created (Jain, 2010). As shown in Figure 7.6, the SSE can not be decreased significantly when the number of clusters is greater than 4. Thus, the optimal number of clusters is 4. Individuals were grouped into 4 categories as shown in Table 7.1. Most of users were grouped into Cluster 1 with relatively low PageRank and a small



Rank	PageRank Score	Screen Name	Community ID	Location	Gender	Followers	Followees	Weibo	Reposts	Comments	Verified Type	Weiage
1	0.12224	Mayili	2	Beijing	Female	26312002	176	4405	1260698	1214021	Hall of Fame	5
2	0.10395	Gillpumpkin	1	Shanghai	Female	2948	106	522	6788	1268	Unverified	3
3	0.04223	Songdandan	3	Beijing	Female	15364225	89	584	2369323	1432161	Hall of Fame	4
4	0.04070	Best Short Jokes	5	Guangdong	Female	10378291	1110	64338	5661735	1186440	Unverified	4
5	0.02113	Labixiaoqiu	2	Guangxi	Female	250005	398	2043	51866	41346	Unverified	5
6	0.01454	Chenqiaoen	1	Taiwan	Female	15485455	414	3150	1353515	590029	Hall of Fame	4
7	0.01205	Yuanjingbo	4	Jiangsu	Female	1315	1167	1576	1239	3084	Unverified	4
8	0.00985	Shenhongfei	2	Shanghai	Male	2607501	608	3914	115288	68401	Hall of Fame	5
9	0.00903	Santfrank	4	Hong Kong	Male	5349	505	3	1	13	Unverified	4
10	0.00877	Psychological Stories	4	Shanghai	Male	1034592	121	9704	361604	33736	Unverified	4

Figure 7.5: Top 10 influential users

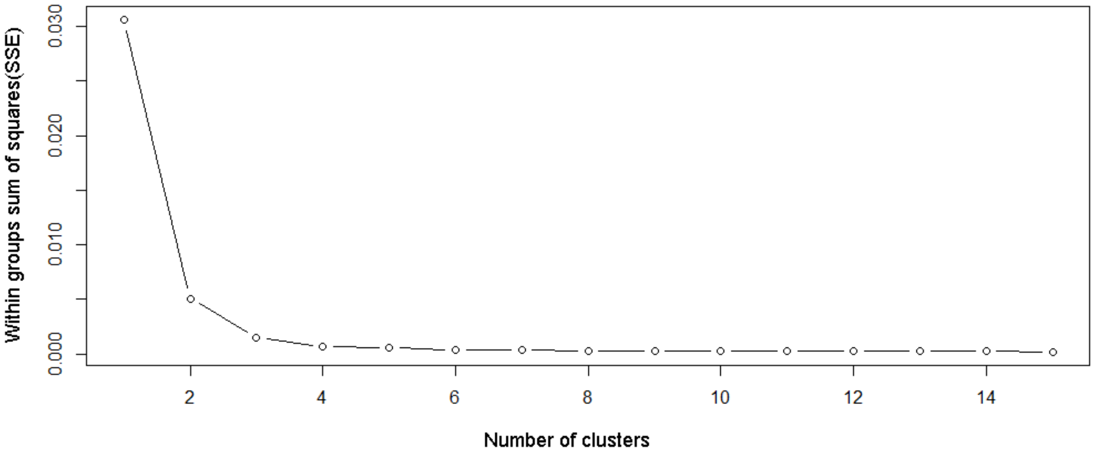


Figure 7.6: Determining the optimal number of clusters by SSE

group of users were grouped into Cluster 2, 3 and 4 with relatively high PageRank. Those groups were weighted in the order of their medoids. Cluster 1 has a weight of $\frac{1}{1+2+3+4}$ (i.e 0.1), Cluster 2 has a weight of $\frac{2}{1+2+3+4}$ (i.e 0.2), Cluster 3 has a weight of $\frac{3}{1+2+3+4}$ (i.e 0.3) and Cluster 4 has a weight of $\frac{4}{1+2+3+4}$ (i.e 0.4). It is worth mentioning that the sum of weights is equal to 1. Then, the weighted average number of reposts of a community is calculated as below:

$$A_{c_i}(\bar{reposts}) = \frac{\sum_{v_k \in c_i} \alpha_{v_k} \times A_{v_k}(reposts)}{N_{c_i}} \tag{7.3}$$

where c_i represents a community, v_k is a vertex in c_i, α_{v_k} is the weight of v_k and N_{c_i} is the number of vertices in c_i. The weighted average number of comments of a community is calculated in the exactly same way. The weight of the edge, $\langle c_i, c_j \rangle$ is calculated by Equation 7.1, where v_i and v_j are replaced with c_i and c_j respectively, $A_{v_i}(reposts)$ and $A_{v_i}(comments)$ are replaced with $A_{c_i}(\bar{reposts})$ and $A_{c_i}(\bar{comments})$ respectively, and $A_{v_i}(social\_diversity)$ is replaced with $A_{c_i}(social\_diversity)$ which is the number of inter-connections c_i initiates. In order to calculate $A_{c_i}(social\_diversity)$, LCC2A-1 was abstracted to LCC2A-2 (see Figure 7.7), in which a group of vertices was contracted into a single vertex (i.e. society). Then, the connections c_i initiated to vertices that are not in the same society with c_i were counted as

$A_{c_i}(social\_diversity)$. The influence of a community was then calculated by Equation 7.2. The proposed weighted average influence takes the difference in individuals' influence into account, where influential users contribute more to the community they belong to.

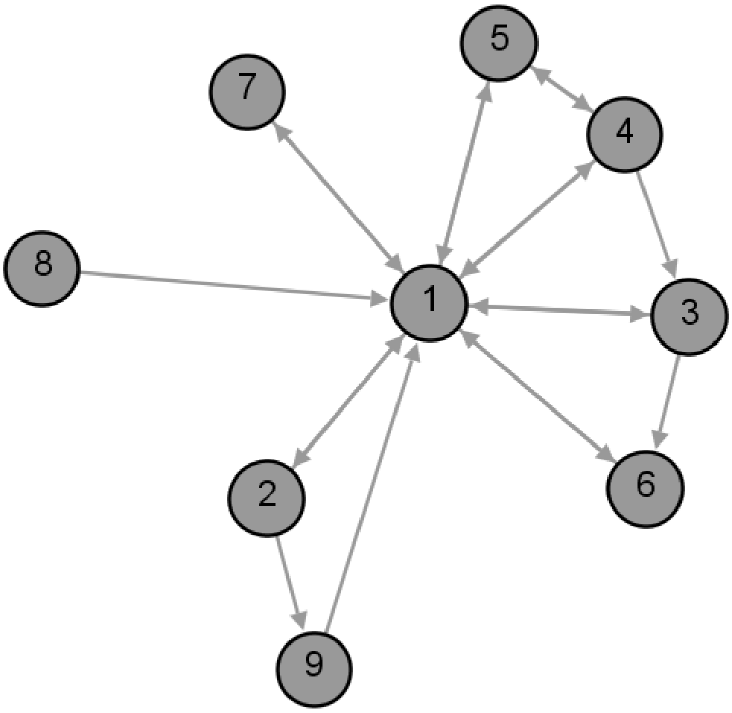


Figure 7.7: LCC2A-2 is a directed graph with 9 vertices and 12 edges. Each vertex in LCC2A-1 represents a society (i.e. community of communities).

Table 7.1: Users are clustered into 4 groups based on their PageRank

Cluster ID	**Size**	**Medoids**
1	4956	4.581472e-05
2	648	3.054315e-04
3	23	6.173495e-03
4	2	1.222421e-01

The top 10 influential communities are listed in Table 7.2, and sorted in descending order according to their PageRank.

Compared to Figure 7.5, which shows that the most influential users come from Community 1, 2, 3, 4 and 5, Table 7.2 shows that the most influential users do not

Table 7.2: The top 10 influential communities

Rank	**Community ID**	**PageRank**	**Size**	**Social diversity**	**Average number of reposts**	**Average number of comments**
1	1	0.10740016	320	5	1562.18	1526.79
2	2	0.09048580	904	7	2538.22	675.53
3	3	0.08581774	746	7	1123.29	663.05
4	8	0.03533766	138	4	637.02	199.87
5	5	0.03498207	58	4	560.89	424.30
6	6	0.03176220	84	3	2937.35	1085.62
7	4	0.02565694	40	3	330.25	314.70
8	19	0.01843035	44	3	40.70	28.52
9	7	0.01520313	154	1	1578.39	1555.29
10	10	0.01437122	72	3	285.90	92.97

all come from the top 5 communities (i.e. Community 1, 2, 3, 8, and 5, sorted in descending order according to their PageRank). Also, Community 6, 7, 8, 10 and 19 do not contain any of the top 10 users (see Figure 7.5) but they are still recognised as the most influential communities.

It is evident that remarkable users do have a substantial impact on the influence of the whole community. For example, as shown in Table 7.2, Community 5 is very modest, compared with Communities 6 and 7, in terms of the size, the social diversity, and the average number of reposts and comments, but it still the 5th most influential community, because one of the most influential users comes from it (i.e. ‘Best Short Jokes’ with 10,378,291 followers, 5,661,735 reposts and 1,186,440 comments).

Table 7.3: The number of users in each cluster. Users of Cluster 4 have the highest PageRank and users in Cluster 1 have the lowest PageRank.

Community ID	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster** 4
4	34	1	5	0
8	131	0	1	0

It is also evident that, however, the influence of a community is not determined alone by remarkable users. For example, as shown in Table 7.3, Community 8 has only 1 user that was categorised into Cluster 3, accounting for 0.76% of the users in the community (i.e. few remarkable users), compared with Community 4 which has 5 users that were categorised into Cluster 3, accounting for 12.50% of the users in the community (i.e. many remarkable users), but Community 8 has a much higher rank than Community 4. Thus, remarkable users alone can contribute significantly to but not completely determine the influence of the communities they belong to.

It was found that 5 of the top 10 communities (i.e. Communities 2, 3, 5, 7 and 10) are incident to Community 8, in which the influence of Communities 2, 3, 5, 7 and 10 are passed on to Community 8. It partially explains why Community 8 is still the 4th ranked community even though it has only 637.02612 reposts and 199.87313 comments on average. Thus, the topological structure is also vital to the dissemination of the influence.

7.4 Potential Applications for Commercial Use

Based on the above findings, a potential application is discussed in this section. Instead of focusing on individual influencers, SNAHOC brings a new perspective on social media marketing, that is making inter-connections with different communities. For example, promoting a newly opened restaurant on Facebook. Instead of pinning the hopes on a local influencer who has the most followers or the loudest mouth, marketers can use SNAHOC to identify the most influential local community and promote the restaurant to the members in the community. Those members can help the marketers to create multiple touch points across different communities very quickly because SNAHOC measures the influence of a community taking the social diversity of the community into account. Although the profile of the selected community is not the loudest, the members of the community have strong connections to the others in different communities, which helps the marketers to disseminate the information rapidly.

Chapter 8

Conclusion

In this research work, a novel approach for social network analysis has been introduced. This is called SNAHOC. This model was designed to explore the hidden information and patterns in social networks along with the levels of abstraction. SNAHOC defines a hierarchy of communities, in which each level is an abstraction of the previous one. Social networks are modelled at multiple levels and information is aggregated every time passing the networks to a higher level of abstraction. Different levels reveal different kinds of information that are not obtainable through classic approaches in which social networks are assumed to be flat.

8.1 Research Contributions

SNAHOC has been thoroughly explained by using graph theory in which the concepts involved in SNAHOC were properly defined and illustrated. SNAHOC is based on a hierarchy of communities, in which each vertex at one level represents a community identified at a previous level. SNAHOC pipes social networks in filters in which information is aggregated in a community scale and pipes them out with a higher level of abstraction.

SinaData, is the data set used in this research work. It was collected from Sina Weibo, through a crawler that was designed for use in this research work.

A case study was conducted to investigate the geographies of communities. By using SNAHOC, a network sampled from SinaData was analysed at multiple levels in which the geographic property of the network was explored at user level, community level and society level. The results of the study have been presented in a visual format to assist with the explanation of the hierarchical nature of the work.

A measure of the influence of a community has been introduced. The influence of a community is based on the influence of users that are within the community and the topological structure the community has. The influence of a user is amplified by the user's social diversity (i.e. the number of inter-connections the user has). Thus, SNAHOC was used to identify the topological structure users have as as well as the topological structure communities have.

8.2 Conclusion

SNAHOC provides a novel way to analyse social networks with different levels of abstraction. Through the demonstration based on real data, SNAHOC has successfully transformed huge amounts of detailed information into summarised information and allowed the exploration of hidden patterns that are not visible at the detailed level.

This thesis gives two case studies. The first case investigated the correlation between the geographic diversity of communities and the number of inter-connections communities have. The number of inter-connections of a community has a positive correlation with the geographic diversity of the community. The geographic diversity of communities is determined by both the richness and evenness of communities in terms of locations. The correlation between the the richness and the number of inter-connections is much stronger than that between the evenness and the number of inter-connections. The second case measured the influence of communities based on a weighted means method. Although remarkable individual users have substantial impacts on the influence of the community they belong to, the influence of communities is not determined by them alone. The topological structure of communities is also vital to the dissemination of the influence.

8.3 Future Work

SNAHOC has a pipeline design that makes it versatile to be adapted in variety of applications. So far, the geographies of communities and the influence of communities have been explored. Other applications, such as identifying trending topics of communities and making a cross comparison with the geographies of communities can tell more about geopolitics in social networks, can also be done using SNAHOC.

As SNAHOC has the ability to create abstractions for social networks, so that social networks can be analysed at a high level of abstraction. This lead to significant reduction in computational time and memory usage. The used abstraction can tell the most about the social networks as unimportant details are filtered out using the abstraction technique. By using SNAHOC, comparison between multiple social networks with huge amount of data will become a more practical proposition.

References

Agarwal, N., Liu, H., Tang, L. and Yu, P. S. (2008). Identifying the influential bloggers in a community. In: *Proceedings of the 2008 international conference on web search and data mining*. ACM, pp. 207–218.

Alef, D. (2010). *Mark Zuckerberg: The face behind Facebook and social networking*. Titans of Fortune Publishing.

Alexa (2014). Alexa traffic rank for weibo.com. URL `http://www.alexa.com/siteinfo/www.weibo.com`.

Andersen, R. and Lang, K. J. (2006). Communities from seed sets. In: *Proceedings of the 15th international conference on World Wide Web*. ACM, pp. 223–232.

Asur, S., Parthasarathy, S. and Ucar, D. (2009). An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **3**(4), p. 16.

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, pp. 44–54.

Backstrom, L., Sun, E. and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 61–70.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, **286**(5439), pp. 509–512.

Batty, M., Hudson-Smith, A., Neuhaus, F. and Gray, S. (2012). Geographic analysis of social network data.

BBC (2014). Facebook emotion experiment sparks criticism. URL `http://www.bbc.com/news/technology-28051930`. 2014-06-30 [Online] (Accessed: 2014-06-30).

Becchetti, L., Boldi, P., Castillo, C. and Gionis, A. (2008). Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 16–24.

Benevenuto, F., Rodrigues, T., Cha, M. and Almeida, V. (2009). Characterizing user behavior in online social networks. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, pp. 49–62.

Bishop, B. (2011). Inside Sina Weibo. *Business Insider*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), p. P10008.

Branigan, T. (2009). China blocks Twitter, Flickr and Hotmail ahead of Tiananmen anniversary. *The Guardian*.

Breese, J. S., Heckerman, D. and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In: *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 43–52.

Cafieri, S., Hansen, P. and Liberti, L. (2014). Improving heuristics for network modularity maximization using an exact algorithm. *Discrete Applied Mathematics*, **163**, pp. 65–72.

Catanese, S., De Meo, P., Ferrara, E. and Fiumara, G. (2010). Analyzing the Facebook friendship graph. *arXiv preprint arXiv:1011.5168*.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys (CSUR)*, **38**(1), p. 2.

Chau, D. H., Pandit, S., Wang, S. and Faloutsos, C. (2007). Parallel crawling for online social networks. In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 1283–1284.

Chen, J. and She, J. (2012). An analysis of verifications in microblogging social networks–Sina Weibo. In: *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on*. IEEE, pp. 147–154.

Clauset, A., Newman, M. E. and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, **70**(6), p. 066111.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C. *et al.* (2001). *Introduction to algorithms*, vol. 2. MIT press Cambridge.

Cui, X., Shi, H. and Yi, X. (2014). Application of association rule mining theory in sina weibo. *Journal of Computer and Communications*, **1**(2), pp. 19–26.

Desikan, P. K. and Srivastava, J. (2008). I/o efficient computation of first order markov measures for large and evolving graphs. In: *Proceedings of WebKDD*. vol. 2008.

Ding, Y., Yan, E., Frazho, A. and Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, **60**(11), pp. 2229–2243.

Dourisboure, Y., Geraci, F. and Pellegrini, M. (2007). Extraction and classification of dense communities in the web. In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 461–470.

Ellison, N. B. *et al.* (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, **13**(1), pp. 210–230.

Erdös, P. and Rényi, A. (1959). On random graphs i. *Publ. Math. Debrecen*, **6**, pp. 290–297.

Flake, G. W., Lawrence, S. and Giles, C. L. (2000). Efficient identification of web communities. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 150–160.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**(3), pp. 75–174.

Fraser, M. and Dutta, S. (2008). *Throwing sheep in the boardroom: How online social networking will transform your life, work and world*. Wiley.

Fuchs, C. (2013). *Social media: A critical introduction*. SAGE Publications.

Gibson, D., Kumar, R. and Tomkins, A. (2005). Discovering large dense subgraphs in massive graphs. In: *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, pp. 721–732.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99**(12), pp. 7821–7826.

Gjoka, M., Kurant, M., Butts, C. T. and Markopoulou, A. (2010). Walking in Facebook: A case study of unbiased sampling of osns. In: *INFOCOM, 2010 Proceedings IEEE*. IEEE, pp. 1–9.

Google (2015). Google Map. URL `https://maps.google.com.au/`.

Grünwald, P. D., Myung, I. J. and Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications.* MIT press.

Gyarmati, L. and Trinh, T. A. (2010). Measuring user behavior in online social networks. *Network, IEEE*, **24**(5), pp. 26–31.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, **15**(4), pp. 784–796.

Hechter, M. (1988). *Principles of group solidarity*, vol. 11. Univ of California Press.

Hui, P., Yoneki, E., Chan, S. Y. and Crowcroft, J. (2007). Distributed community detection in delay tolerant networks. In: *Proceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture.* ACM, p. 7.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, **31**(8), pp. 651–666.

Java, A., Joshi, A. and Finin, T. (2008). Detecting commmunities via simultaneous clustering of graphs and folksonomies. In: *Proceedings of the Tenth Workshop on Web Mining and Web Usage Analysis (WebKDD).* ACM.

Jehn, K. A., Northcraft, G. B. and Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly*, **44**(4), pp. 741–763.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons.

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, pp. 137–146.

Kirkpatrick, D. (2010). *The Facebook effect: The inside story of the company that is connecting the world*. Simon & Schuster.

Knoke, D. and Yang, S. (2008). *Social network analysis*. No. no. 154 in Quantitative Applications in the Social Sciences. SAGE Publications.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010). What is Twitter, a social network or a news media? In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 591–600.

Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 631–636.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, **58**(7), pp. 1019–1031.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P. and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, pp. 29–42.

Molles, M. C. and Cahill, J. F. (1999). *Ecology: concepts and applications*. WCB/McGraw-Hill Dubuque, IA.

Moore, M. (2011). Top five foreigners using Sina Weibo. URL `http://www.telegraph.co.uk/technology/8736949/Top-five-foreigners-using-Sina-Weibo.html`.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *science*, **328**(5980), pp. 876–878.

Mulder, C., Bazeley-White, E., Dimitrakopoulos, P., Hector, A., Scherer-Lorenzen, M. and Schmid, B. (2004). Species evenness and productivity in experimental plant communities. *Oikos*, **107**(1), pp. 50–63.

Newman, M. E. (2004). Analysis of weighted networks. *Physical Review E*, **70**(5), p. 056131.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, **69**(2), p. 026113.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Palla, G., Barabási, A.-L. and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, **446**(7136), pp. 664–667.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In: *Computer and Information Sciences-ISCIS 2005*, Springer, pp. 284–293.

Raghavan, U. N., Albert, R. and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, **76**(3), p. 036106.

Reagans, R. and Zuckerman, E. W. (2001). Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science*, **12**(4), pp. 502–517.

Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 61–70.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, **14**(5), pp. 465–471.

Rivlin, G. (2006). Wallflower at the web party. *New York Times.*

Ronhovde, P. and Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, **80**(1), p. 016109.

Rosenfeld, J. M. (2002). Spiders and crawlers and bots, oh my: The economic efficiency and public policy of online contracts that restrict data collection. *Stan. Tech. L. Rev.*, **2002**, pp. 3–4.

Rosvall, M. and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, **104**(18), pp. 7327–7331.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105**(4), pp. 1118–1123.

Scellato, S., Mascolo, C., Musolesi, M. and Latora, V. (2010). Distance matters: geo-social metrics for online social networks. In: *Proceedings of the 3rd conference on Online social networks*. pp. 8–8.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**(1), pp. 3–55.

Sina (2014). Sina Weibo: Second quarter 2014 financial results.

Sun, Y., Danila, B., Josić, K. and Bassler, K. E. (2009). Improved community structure detection using a modified fine-tuning strategy. *EPL (Europhysics Letters)*, **86**(2), p. 28004.

Tang, L., Liu, H. and Zhang, J. (2012). Identifying evolving groups in dynamic multimode networks. *Knowledge and Data Engineering, IEEE Transactions on*, **24**(1), pp. 72–85.

Tang, L., Liu, H., Zhang, J. and Nazeri, Z. (2008). Community evolution in dynamic multi-mode networks. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 677–685.

Tang, L., Wang, X. and Liu, H. (2009). Uncoverning groups via heterogeneous interaction analysis. In: *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, pp. 503–512.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, pp. 425–443.

Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. (2011). The anatomy of the Facebook social graph. *arXiv preprint arXiv:1111.4503*.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, vol. 8. Cambridge university press.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *nature*, **393**(6684), pp. 440–442.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. and Zhao, B. Y. (2009). User interactions in social networks and their implications. In: *Proceedings of the 4th ACM European conference on Computer systems*. Acm, pp. 205–218.

Wu, S., Hofman, J. M., Mason, W. A. and Watts, D. J. (2011). Who says what to whom on Twitter. In: *Proceedings of the 20th international conference on World wide web*. ACM, pp. 705–714.

Yan, Q., Wu, L. and Zheng, L. (2013). Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and Its Applications*, **392**(7), pp. 1712–1723.

Ye, S., Lang, J. and Wu, F. (2010). Crawling online social graphs. In: *Web Conference (APWEB), 2010 12th International Asia-Pacific*. IEEE, pp. 236–242.

Yu, P. S., Li, X. and Liu, B. (2004). On the temporal dimension of search. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, pp. 448–449.

Zeng, H.-J., Chen, Z. and Ma, W.-Y. (2002). A unified framework for clustering heterogeneous web objects. In: *Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on*. IEEE, pp. 161–170.

Appendix A

Journal of Computer and Communications, **2014, 2, 19-26**
Published Online January 2014 (http://www.scirp.org/journal/jcc)
http://dx.doi.org/10.4236/jcc.2014.21004

Application of Association Rule Mining Theory in Sina Weibo

Xiao Cui, Hao Shi, Xun Yi

College of Engineering and Science, Victoria University, Melbourne, Australia.
Email: xiao.cui1@live.vu.edu.au

Received November 21st, 2013; revised December 17th, 2013; accepted December 24th, 2013

ABSTRACT

A user profile contains information about a user. A substantial effort has been made so as to understand users' behavior through analyzing their profile data. Online social networks provide an enormous amount of such information for researchers. Sina Weibo, a Twitter-like microblogging platform, has achieved a great success in China although studies on it are still in an initial state. This paper aims to explore the relationships among different profile attributes in Sina Weibo. We use the techniques of association rule mining to identify the dependency among the attributes and we found that if a user's posts are welcomed, he or she is more likely to have a large number of followers. Our results demonstrate how the relationships among the profile attributes are affected by a user's verified type. We also put some efforts on data transformation and analyze the influence of the statistical properties of the data distribution on data discretization.

KEYWORDS

Association Rules; User Profiles; Sina Weibo; Social Network

1. Introduction

Online social networks such as Facebook, Twitter and Google+ have become an integral part of people's daily lives. No matter how they differentiate from one another, user profiles are a key feature. A user profile may include but not be limited to gender, age, location, occupation, social contacts, etc. The availability of the information may vary from one site to another. In spite of the fact that user profiles are less dynamic than other online behaviors, they still provide a clear signal of users' characteristics. A substantial effort has been made recently in order to obtain knowledge about users from their profile data. Lampe *et al.* [1] found that profile completion percentage on Facebook has a positive relationship with the number of friends a user has. Mislove *et al.* [2] proposed an algorithm to infer the missing part of a user profile according to other similar profiles. Quercia *et al.* [3] conducted a study on the relationship between the Big Five personality traits and user behaviors on Twitter. They introduced a novel method to predict the personality based on the number of followers, followings and tweets a user has.

As Twitter is banned in China, Sina Weibo is considered a replacement for it. Sina Weibo has reached 56 million daily active users (who spend an average of one hour per day with the service) [4]. Sina Weibo has had a significant influence on Chinese society. Unlike its predecessors, studies on Sina Weibo are still in an initial state. There are a few studies on Sina Weibo with regard to user profiles. Guo *et al.* [5] found that the connections between users are mostly one-way and the number of followers a user has changes very fast. Chen and She [6] carried out a similar study but compared verified users with unverified ones. They believed that users whose real identity has been verified are more likely to have greater influence. Wang *et al.* [7] examined the correlation between the number of followers, followings and posts. They found that the number of followers grows rapidly as the number of followings increases from 10 to 3000. They also stated that the increase in posts can lead to more followers as long as the number of posts does not exceed 20,000.

Although considerable attention has been paid to Sina Weibo, associations among different profile attributes, such as the association between the number of reposts and comments, have not been well examined yet. Due to the fact that a large number of users on Sina Weibo have been verified according to their professional background, people on Sina Weibo are more likely to act responsibly and engage honestly with the community. It is worthwhile to explore users' characteristics on Sina Weibo especially considering they have different verified types (e.g. local authorities, news agency, and celebrity). Our research is based on a set of first-hand data collected from Sina Weibo, containing 1,192,972 users' profiles. The major contributions are summarized as follows:

- Continuous data (e.g. the number of followers) are replaced by meaningful labels (e.g. the grass roots and social star).
- The influence of the distribution of the data on data discretization is analyzed.
- Association rule mining is conducted with respect to users' verified types.
- A comparison between different types of users is made.

The rest of the paper is organized as follows. Section 2 presents the data model used in this paper. Definitions such as the number of followings a user has are given. Section 3 explains the process of data collection. The social relationships among users in Sina Weibo are illustrated. Section 4 discusses the methods for data discretization. The statistical properties of the data distribution are taken into consideration. Section 5 introduces an Apriori-based method for association rule mining and explains how we are going to conduct the association rule mining in Sina Weibo. Empirical results are given in Section 6 and conclusions are drawn in the last section.

2. Data Model

The information in a user profile may include various attributes of a user such as geographical location, academic and professional background, interests, preferences, etc. The availability of such information may vary from one site to another. In terms of microblogging, *i.e.* Sina Weibo, the number of followers, followings and posts a user has are three indispensable parts of a user profile. Such information is always displayed at a prominent place. Besides, a verified type is added to a user profile as users on Sina Weibo may choose to verify their identity based on their professional background. In this paper, a user profile is defined as follows:

profile(*uid*) = {*username*, *province*, *gender*, *number of followers*, *number of followings*, *number of posts*, *number of reposts*, *number of comments*, *verified type*, *time since created*}

Each user has a unique identification number (*uid*). The core attributes of a profile are defined as follows:

- *NoA* refers to *number of followers*. *NoA*(*uid*) is the total number of audience who are listening to the broadcast of user *uid*. *NoA* is one of the major signs of a user's popularity.
- *NoB* refers to *number of followings*. *NoB*(*uid*) is the total number of broadcasts to which user *uid* is listening. Sina Weibo enforces that a user can listen a maximum of 2000 broadcasts.
- *NoP* refers to *number of posts*. *NoP*(*uid*) is the total number of posts that user *uid* updates. *NoP* can be a good indicator of a user's activeness.
- *NoR* refers to *number of reposts*. *NoR*(*uid*) is the total number of reposts that others forward from user *uid*. *NoR* is a sign of the capability a user has to spread out the information.
- *NoC* refers to *number of comments*. *NoC*(*uid*) is the total number of comments others leave on user *uid*. *NoC* can reveal the likelihood of a user to initiate a hot topic.
- *VT* refers to *verified type*. *VT* includes *red star* (an ordinary user whose real identity is verified), *beauty*, *e-celebrity*, *corporation*, *government*, *media*, *organization*, *campus*, *application software*, and *website* [8]. For example, user Xinhua News Agency, the official press agency of China, is classed as *media*.
- *TsC* refers to *time since created*.

3. Data Collection

Users' profiles are collected through the REST API provided by Sina Weibo. Bilateral relationships are used to expand the search of new users. Social relationships among users are defined as follows (see Figure 1).

Scenario 1 indicates that *uid*1 and *uid*2 have no connection between them. Scenario 2 shows that *uid*1 is a follower of *uid*2. Scenario 3 explains bilateral friendships where *uid*1 is a follower of *uid*2 and *uid*2 is also a follower of *uid*1. We assume that is two users follow each other, they are considered friends.

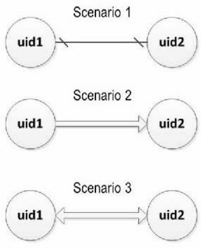


Figure 1. Social relationships.

Getting the friends of a friend is the strategy used in this paper to obtain users' ids from Sina Weibo. The REST API provides facilities to retrieve profile information according to a user's id. The implementation details are given below (see **Table 1**).

Unlike studies [4-6] where a user's followings are used to expand the search of new users, bilateral relationships are used in this study. Users who follow each other seem to have a closer relation between them. This method can prevent the search of new users from the spammers because no one likes to subscribe a spammer's microblog.

Finally, 1,192,972 users' profiles are retrieved. 39.58% of them are verified users. *Red star* and *e-celebrity* account for 91.08% of verified users (see **Figure 2**).

4. Data Discretization

Data mining process involves a preprocessing step in order to assure the data have the quality and the format required by the algorithm. Users are classified by their attributes. For example, according to *NoA*, users are classified into two groups: *the grass roots* and *social star*. Users in the latter group have much more followers than users in the former one. Other continuous data are replaced as well in a similar way (see **Table 2**).

Table 1. Pseudocode for data collection.

```
enqueue i in q
while q is not empty do
    get friends_uids according to i
    for each j in friends_uids do
        if j does not exist in q then
            insert j into q
        end
    end
    dequeue k from q
    get profile according to k
    set i = k
end
```

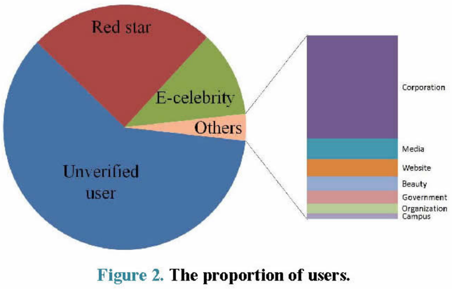


Figure 2. The proportion of users.

4.1. The *K*-Means Method and the Pareto Principle

This paper experimented with two methods: the *k*-means clustering algorithm and the Pareto principle. The purpose of clustering is to search for similar examples and group them into clusters such that the distance between examples within cluster is as small as possible and the distance between clusters is as large as possible [9]. Let $P = \{p_1, p_2, \cdots, p_n\}$ be a set of data points to be clustered and k is the number of clusters (Here, $k = 2$). Randomly select k data points from P as the initial centroids of the clusters, $C = \{c_1, c_2\}$. Then, following steps are repeatedly performed until the convergence is obtained: 1) Assign each data point $p_i \in P, i = \{1, 2, \cdots, n\}$ to the closest centroid either c_1 or c_2. 2) Re-compute the centroids of the clusters $C = \{c_1, c_2\}$. Centroid is the mean of the points in cluster.

The Pareto principle (also known as the 80 - 20 rule) [10] originally referred to the observation that 80% of Italy's wealth belonged to only 20% of the population. Here, we assume that, for example, a user whose followers are more than 80% of the other users is classed as *social star*. The quantile function used to calculate the cut points between the groups (e.g. *the grass roots* and *social star*), is defined as follows [11]:

$$F^{-1}(p) = \min\{X \in \mathbb{R} : F(X) \geq p\},\ p = 0.8 \qquad (1)$$

Here, X may refer to one of the variables in **Table 2**. The distribution function of X is given by $F(X) = P(X \leq x)$ where $F(X)$ represents the probability that X is less than or equal to x. Equation (1) determines the place where 80% of the data lies below it, e.g. 80% of *NoA* is less than or equal to 1140% and 80% of *NoR* is less than or equal to 294.

4.2. Discretization Index

In this paper, a discretization index (di) is proposed to measure the quality of the discretization produced from above methods. Let $X = \{x_1, x_1, \cdots, x_n\}$ be a set of data points to be split. Suppose X is partitioned into two groups $G = \{g_1, g_2\}$. A di is defined as follows:

$$di = \max_{i \in \{1,2\}}(\delta_i) \sum_{j=1}^{2} \sum_{x_k \in g_j} (x_k - \mu_j) \qquad (2)$$

where δ_i denotes the proportion of g_i in X and μ_j denotes the mean of data points in g_j. The method with the smallest di is considered the best method based on the following criteria: 1) Minimize the distances within the clusters and maximize the distances between the clusters. 2) Split as equally as possible. The reason why both criteria are needed is that using the first criterion (*i.e.*, clusters that are coherent internally but clearly different from each other) alone to split the data may cause an extremely uneven partition (see Section 4.3). As asso-

Table 2. Data discretization.

Continuous Data	Statistical Properties		Partition			di (10^9)	
	Standard deviation	*Skewness*	*Class*	*Quantity*[a]	*Interval*[b]	*K-means*	*Pareto*
NoA	219583.30	116.64	The grass roots	1,192,865	[1.1140)	15.31	12.67
			Social star	107	[1140.63717128]		
NoB	485.80	2.00	Self-centered	1,029,534	[1.894)	0.19	8.06
			Scout	163,438	[894.2000]		
NoP	3032.53	12.25	Lurker	1,152,976	[1.2034)	12.36	0.88
			Blog zealot	39,996	[2034.413549]		
NoR	62489.64	131.02	Valuelss	1,192,897	[0.294)	4.10	3.44
			Propagator	75	[294.18645439]		
NoC	56012.94	353.34	Uninterested	1,192,939	[0.206)	3.05	2.22
			Topic inititator	33	[1206.39344085]		

[a] Calculation was based on the partitions generated from the k-means method; [b] Calculation was based on the partitions generated from the Pareto principle.

ciation rules are generated from frequent itemsets (see Section 5), data in the minority, for example, 200 *social star* users in 1,192,972 users, are very likely to be overlooked. More explanations for why partitioning as evenly as possible is important to this study are given in Section 5. We propose di aiming to build a balance between the criteria.

4.3. Comparison between the Methods

We found that the use of discretization methods depends on the statistical properties of the data distribution. The spread of the data (*i.e.* standard deviation) and the symmetry of the data (*i.e.* skewness) may have significant influence on the performance of the discretization. Higher standard deviation implies greater spread of data. Positive values for the skewness indicate that the distribution is skewed right. Higher skewness implies longer tail in the right side. A normal distribution has a skewness of 0. We found that the *k*-means method is very good at creating clusters coherent internally but different from each other. However, the *k*-means method tends to partition data in an extremely uneven way when the distribution is skewed (see **Table 2**). On the other hand, partition based on the Pareto principle produces a lower di in most cases (see **Table 2**). Data are partitioned in a 80-20 way without impairing the internal coherence and the external difference of the clusters.

We use examples to illustrate how the statistical properties of the data distribution can have influence on the data discretization. As shown in **Figure 3**, the distribution of *NoB* is much closer to a normal distribution with a skewness of 2, at the same time it has the lowest standard deviation compared with other variables. In this case, the *k*-means method produces a lower di than the Pareto principle. In comparison, data points in *NoA* are spread out over an extremely large range of values 1 to 63,717,128. A skewness of 116.64 indicates that the distribution of *NoA* has a very long tail at the right side (see **Figure 3**). As a consequence, the majority of data points in *NoA* fall within a very small range of values and very few of data points fall within an extremely wide range of values. Actually, 80% of the data points in *NoA* fall within the interval [1, 1140) and the rest falls within the interval [1140, 63,717,128]. In this case, the *k*-means method tends to group almost all data points into one cluster and put the rest into another one. Actually, only 0.01% of users were classed as *social star* in the *k*-means method (see **Table 2**). Partition based on the Pareto principle is applied in this study because it makes a trade-off between the criteria.

5. Mining Association Rules in Sina Weibo

5.1. Association Rule Mining

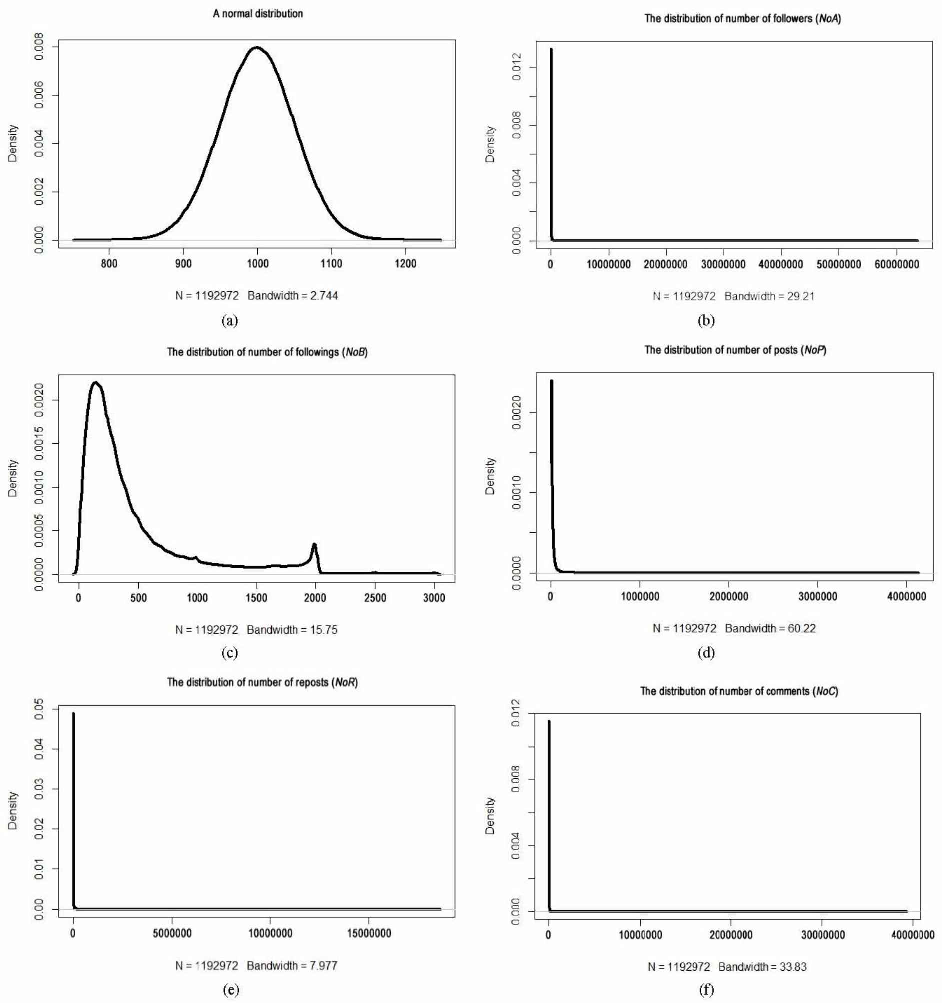
The association rule mining can be conceptualized as follows [9]: Let $f = \{I_1, I_2, \cdots, I_n\}$ be the set of all items. Let DB be a set of database transactions where each transaction T is a set of items such that $T \subseteq f$. Let A be a set of items. A transaction is said to contain A if and only if $A \subseteq f$. An association rule is an implication of the form $A \Rightarrow B[s, c, l]$, where $A \subseteq f$, $B \subseteq f$, $A \cap B = \varnothing$. The support s, confidence c and lift l of the rule $A \Rightarrow B$ are defined as:

$$s = P(A \cup B) = F(A \cup B) / |DB| \tag{2}$$

$$c = P(B|A) = F(A \cup B) / F(A) \tag{3}$$

$$l = P(A \cup B) / P(A)P(B) \tag{4}$$

where $F(A)$ stands for the number of transactions containing the set X in DB and $|DB|$ denotes the total number of transactions in DB. Rules with the support more than a minimum support threshold $s_{\min}$ and the confidence more than a minimum confidence threshold $c_{\min}$ are called strong. A set of items is refe-



(a) (b) (c) (d) (e) (f)

Figure 3. Data distribution.

reed as an itemset. An itemset that contains (*k*) items is a *k*-itemset. The support count of an itemset is the number of transactions containing the itemset. The minimum support count is defined as $s_{\min} \cdot |DB|$. An itemset is frequent if its support count is not less than the minimum support count.

5.2. Apriori Algorithm

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses the Apriori property, *i.e.*, all nonempty subsets of a frequent itemset much also be frequent. Let L_i be the set of frequent i-itemsets. Given L_{k-1}, Apriori algorithm finds L_k using join and prune actions as follows: 1) Join: To find L_k, a set of candidate *k*-itemsets, denoted as l_k is generated by joining L_{k-1} with itself. Any two (*k* − 1)-itemsets A and B are joinable if they contain (*k* −

2) common items. For example, $A=\{x_1,\cdots,x_{(k-2)},x_{(k-1)}\}$ and $B=\{x_1,\cdots,x_{(k-2)},x_k\}$ are joinable. The resulting candidate k-itemset is $\{x_1,\cdots,x_{(k-2)},x_{(k-1)},x_k\}$. 2) Prune: l_k can be huge. To reduce the size of l_k, the Apriori property is used as follows. If any (k − 1)-subset of an candidate k-itemset is not in L_{k-1}, the candidate cannot be frequent either and so can be removed from l_k. The set of remaining candidates in l_k is a superset of L_i, that is, its elements may or may not be frequent, but all of the frequent k-itemsets must be included in l_k. A scan of the database to determine the count of each candidate in l_k would result in the determination of L_k, *i.e.*, all candidate having a count no less than the minimum support count are frequent and therefore belong to L_k. By Apriori algorithm, all frequent itemsets along with their support counts can be found efficiently.

5.3. Experimental Design

Suppose dataset *U* contains all the data collected from Sina Weibo. Association rules are mined from both *U* and its subsets.

Considering the property of association rule mining described above, rare types of users are very likely to be pruned due to their relatively low support counts. Splitting *U* into disjoint subsets based on *VT* and mining association rules from them separately is necessary so as to avoid overlooking some interesting patterns that are hidden in the rare types of users. The dataset *U* is divided into 2 subsets: *verified_accounts* and *unverified_accounts*. A comparison in terms of association rules, between *verified_accounts* and *unverified_accounts*, is made to identify the difference between verified users and unverified users. If necessary, the dataset *verified_accounts* can be further divided according to *VT*. In this paper, a comparison between *red star* and *e-celebrity* is conducted for two reasons: 1) *red star* and *e-celebrity* together account for 91.08% of verified users. 2) *red star* refers to the masses, opposite to *e-celebrity* who are public figures and professionals and well known in local communities.

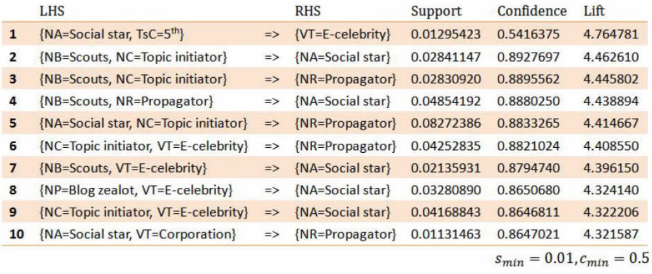
Considering the fact that users who have large number of followers (followings, posts, reposts, and comments) only account for a very small part, we have to give a relatively small $s_{\min}$ so as to assure the rules for *social star* (*scout*, *blog zealot*, *propagator*, and *topic initiator*) can be elicited. Association rules are sorted by lift values. A lift equals to 1 means the occurrence of A is independent of the occurrence of B if an association rule is in the form of $A \Rightarrow B[s,c,l]$. A lift is greater than 1 indicates that the occurrence of A has a positive effect on the occurrence of B. We are interested in the profile attributes which are dependent on each other.

6. Empirical Results

We found that both *NoR* and *NoC* play important roles in a user's popularity (see **Figure 4**). If a user's posts are welcomed, either the posts are forwarded by many times or many people leave comments about the posts, the owner of the posts is much more likely to be tagged as a *social star*. Another finding was that *NoC* is positively correlated with *NoR*. 3 of 4 rules for "*NoR* = Propagator" were attributed to "*NoC* = Topic initiator" (*i.e.* Rule #3, 5, and 6). Also, an *e-celebrity* user is always accompanied by a large number of followers (*social star*). Thus, *social star* is a good indicator of *e-celebrity*. We found that a 5-year *social star* user is an *e-celebrity* with a confidence of 54.16%.

A comparison between association rules derived from *unverified_accounts* and *verified_accounts* was made (see **Figures 5** and **6**).

A positive correlation between *NoC* and *NoR* exists in both of them; however, rules in *unverified_accounts* have higher lift than that in *verified_accounts*. In other words, *NoC* and *NoR* are more dependent on one another in *unverified_accounts*. According to above findings, we could state that, for a verified user, an increase in *NoC* may not enhance the probability of an increase in *NoR*. On the other hand, for an ordinary user who has not been verified yet, saying something controversial to receive more comments (*NoC*) is a good way to increase the rate of



	LHS		RHS	Support	Confidence	Lift
1	{NA=Social star, TsC=5th}	=>	{VT=E-celebrity}	0.01295423	0.5416375	4.764781
2	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.02841147	0.8927697	4.462610
3	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.02830920	0.8895562	4.445802
4	{NB=Scouts, NR=Propagator}	=>	{NA=Social star}	0.04854192	0.8880250	4.438894
5	{NA=Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.08272386	0.8833265	4.414667
6	{NC=Topic initiator, VT=E-celebrity}	=>	{NR=Propagator}	0.04252835	0.8821024	4.408550
7	{NB=Scouts, VT=E-celebrity}	=>	{NA=Social star}	0.02135931	0.8794740	4.396150
8	{NP=Blog zealot, VT=E-celebrity}	=>	{NA=Social star}	0.03280890	0.8650680	4.324140
9	{NC=Topic initiator, VT=E-celebrity}	=>	{NA=Social star}	0.04168843	0.8646811	4.322206
10	{NA=Social star, VT=Corporation}	=>	{NR=Propagator}	0.01131463	0.8647021	4.321587

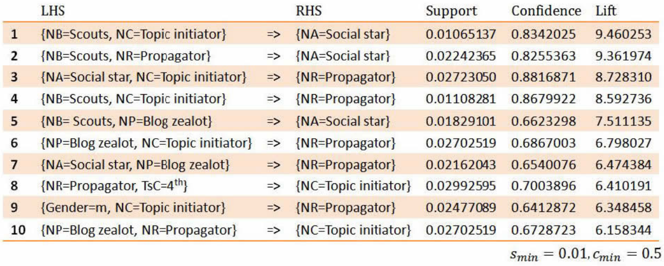
$s_{min} = 0.01, c_{min} = 0.5$

Figure 4. Top 10 rules derived from *U*.

diffusion of his or her posts (*NoR*). Actually, it has already happened in many online social networks where people initiate some controversial topics in order to become famous [12].

Although *red_star* is a disjoint subset of *verified_accounts*, the dependence, in terms of lift, between *NoC* and *NoR* in *red_star* is much stronger than that in *verified_accounts* itself (see **Figure 7**). On the other hand, rules derived from *e-celebrity* are less interesting in terms of lift (see **Figure 8**).

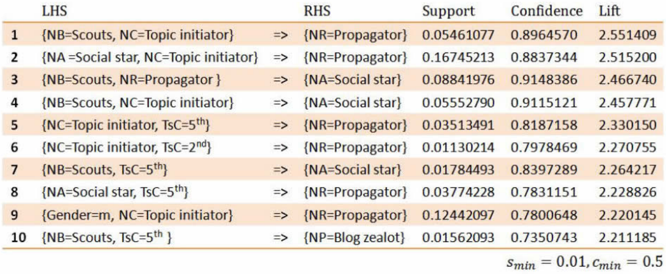
We found that if an *e-celebrity* user's posts are welcomed, then he or she is a *blog zealot* with a confidence greater than 65%. Actually, it happens in many kinds of user types. Unlike *red star* users, users such as *corporation*, *media*, and *application software*, have a strong motive for promoting themselves or something else. As a consequence, they are likely to send as many messages as possible. At the same time, due to their high reputation,



	LHS		RHS	Support	Confidence	Lift
1	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.01065137	0.8342025	9.460253
2	{NB=Scouts, NR=Propagator}	=>	{NA=Social star}	0.02242365	0.8255363	9.361974
3	{NA=Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.02723050	0.8816871	8.728310
4	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.01108281	0.8679922	8.592736
5	{NB= Scouts, NP=Blog zealot}	=>	{NA=Social star}	0.01829101	0.6623298	7.511135
6	{NP=Blog zealot, NC=Topic initiator}	=>	{NR=Propagator}	0.02702519	0.6867003	6.798027
7	{NA=Social star, NP=Blog zealot}	=>	{NR=Propagator}	0.02162043	0.6540076	6.474384
8	{NR=Propagator, TsC=4th}	=>	{NC=Topic initiator}	0.02992595	0.7003896	6.410191
9	{Gender=m, NC=Topic initiator}	=>	{NR=Propagator}	0.02477089	0.6412872	6.348458
10	{NP=Blog zealot, NR=Propagator}	=>	{NC=Topic initiator}	0.02702519	0.6728723	6.158344

$s_{min} = 0.01, c_{min} = 0.5$

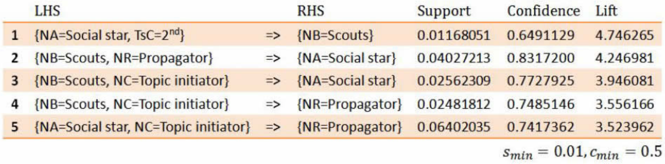
Figure 5. Top 10 rules derived from *unverified_accounts*.



	LHS		RHS	Support	Confidence	Lift
1	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.05461077	0.8964570	2.551409
2	{NA =Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.16745213	0.8837344	2.515200
3	{NB=Scouts, NR=Propagator }	=>	{NA=Social star}	0.08841976	0.9148386	2.466740
4	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.05552790	0.9115121	2.457771
5	{NC=Topic initiator, TsC=5th}	=>	{NR=Propagator}	0.03513491	0.8187158	2.330150
6	{NC=Topic initiator, TsC=2nd}	=>	{NR=Propagator}	0.01130214	0.7978469	2.270755
7	{NB=Scouts, TsC=5th}	=>	{NA=Social star}	0.01784493	0.8397289	2.264217
8	{NA=Social star, TsC=5th}	=>	{NR=Propagator}	0.03774228	0.7831151	2.228826
9	{Gender=m, NC=Topic initiator}	=>	{NR=Propagator}	0.12442097	0.7800648	2.220145
10	{NB=Scouts, TsC=5th }	=>	{NP=Blog zealot}	0.01562093	0.7350743	2.211185

$s_{min} = 0.01, c_{min} = 0.5$

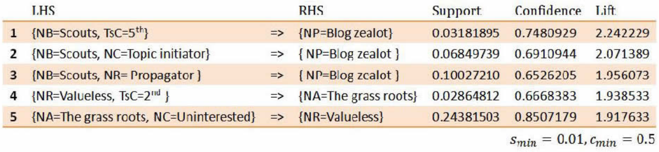
Figure 6. Top 10 rules derived from *verified_accounts*.



	LHS		RHS	Support	Confidence	Lift
1	{NA=Social star, TsC=2nd}	=>	{NB=Scouts}	0.01168051	0.6491129	4.746265
2	{NB=Scouts, NR=Propagator}	=>	{NA=Social star}	0.04027213	0.8317200	4.246981
3	{NB=Scouts, NC=Topic initiator}	=>	{NA=Social star}	0.02562309	0.7727925	3.946081
4	{NB=Scouts, NC=Topic initiator}	=>	{NR=Propagator}	0.02481812	0.7485146	3.556166
5	{NA=Social star, NC=Topic initiator}	=>	{NR=Propagator}	0.06402035	0.7417362	3.523962

$s_{min} = 0.01, c_{min} = 0.5$

Figure 7. Top 5 rules derived from *red_star*.



	LHS		RHS	Support	Confidence	Lift
1	{NB=Scouts, TsC=5th}	=>	{NP=Blog zealot}	0.03181895	0.7480929	2.242229
2	{NB=Scouts, NC=Topic initiator}	=>	{ NP=Blog zealot }	0.06849739	0.6910944	2.071389
3	{NB=Scouts, NR= Propagator }	=>	{ NP=Blog zcalot }	0.10027210	0.6526205	1.956073
4	{NR=Valueless, TsC=2nd }	=>	{NA=The grass roots}	0.02864812	0.6668383	1.938533
5	{NA=The grass roots, NC=Uninterested}	=>	{NR=Valueless}	0.24381503	0.8507179	1.917633

$s_{min} = 0.01, c_{min} = 0.5$

Figure 8. Top 5 rules derived from *e-celebrity*.

other users prefer to forward their posts or have discussion with them. Posts are welcomed is independent of having a large number of posts. For this reason, lift values in **Figure 8** are very close to 1.

7. Conclusion

In this study, we explored the relationships among different profile attributes through the techniques of association rule mining. We found that a user is more likely to have a large number of followers (*NoA*) if his or her posts are forwarded by many times (*NoR*) or many people get involved in the discussion he or she initiated (*NoC*). Our results indicate that *NoR* and *NoC* are strongly dependent on each other with respect to ordinary users (both unverified users and *red star* users). Profile attributes for verified users are relatively independent on one another. We also examined both the *k*-means method and the Pareto principle as a method for data discretization. We found that the statistical properties of data distribution can have significant influence on data discretization. Due to the fact that data used in this study are skewed heavily, we suggested using the Pareto principle to partition data.

REFERENCES

[1] C. A. Lampe, N. Ellison and C. Steinfield, "A Familiar Face (Book): Profile Elements as Signals in an Online Social Network," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 435-444.

[2] A. Mislove, B. Viswanath, K. P. Gummadi and P. Druschel, "You Are Who You Know: Inferring User Profiles in Online Social Networks," *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010, pp. 251-260.

[3] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011 *IEEE Third International Conference on Privacy, Security, Risk and Trust* (*Passat*) *and* 2011 *IEEE Third International Conference on Social Computing* (*Socialcom*), 2011, pp. 180-185.

[4] D. Clark, R. Crandall and Y. Mei, "4th Annual China 2.0 Conference Underscores Business Innovation, Social Impact and U.S.-China Links," 2013. http://sprie.gsb.stanford.edu/news/4th_annual_china_20_confe-rence_underscores_business_innovation_social_impact_and_uschina_links_20131022/

[5] Z. Guo, Z. Li, H. Tu and L. Li, "Characterizing User Behavior in Weibo," 2012 *Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing* (*MUSIC*), 2012, pp. 60-65.

[6] J. Chen and J She, "An Analysis of Verifications in Microblogging Social Networks—Sina Weibo," 2012 32*nd International Conference on Distributed Computing Systems Workshops* (*ICDCSW*), 2012, pp. 147-154.

[7] C. Wang, X. Guan, T. Qin and W. Li, "Who Are Active? An In-Depth Measurement on User Activity Characteristics in Sina Microblogging," *Global Communications Conference* (*GLOBECOM*), 2012, pp. 2083-2088.

[8] Sina Open API. http://open.weibo.com/wiki/

[9] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2006.

[10] J. M. Juran and A. B. Godfrey, "Juran's Quality Handbook (Vol.2)," McGraw Hill, New York, 1999.

[11] I. Frohne and R. J. Hyndman, "Sample Quantiles," R Project, 2009.

[12] J. Feng, "Romancing the Internet: Producing and Consuming Chinese Web Romance," Brill, 2013. http://dx.doi.org/10.1163/9789004259720