Growth Dynamics of 'Socl': an Interest-Based Social Network

By

Aisha Al Zbeidi

A Thesis Presented to the

Masdar Institute of Science and Technology

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

In

Computing and Information Science

Growth Dynamics of 'Socl': an Interest-based Social Network

By Aisha Al Zbeidi

A Thesis Presented to the Masdar Institute of Science and Technology in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Computing and Information Science

May 2014

AUTHOR'S DECLARATION

I understand that copyright in my thesis is transferred to Masdar Institute of Science and Technology.

Author Aisha Al Zbeidi

RESEARCH SUPERVISORY COMMITTEE MEMBERS

Dr. Iyad Rahwan, _____

Masdar Institute of Science and Technology

Dr. Wei Lee Woon, _____

Masdar Institute of Science and Technology

Dr. Zeyar Aung, _____

Masdar Institute of Science and Technology

Abstract

Online Social Networks have been attracting millions of users who have adopted these sites in their daily lives activities. As the amount of information and the real-time data uploaded to these systems increasing, it becomes a rich source for researchers from industry and academic institutions to study the human behavior patterns for different purposes. The structure of these expanding networks is expected to mirror the structure of real-life communications and relationships in society.

In this thesis, we adopted graph theory models and some statistical properties and measurements of the social networks to analyze the dynamic growth in 'Socl' which is an interest-based social network developed by Microsoft FUSE Labs. We studied the changes in the properties of the users' friendship network quantified based on their interactions and we presented a series of measurements of the users' friendship and interests behavior in the range of time between January and November 2012.

We observed interesting evolutionary patterns on the topological structure of the 'Socl' network as it went from a private project within Microsoft Corporation to a shared project with students from three universities in an invitation-only stage before it released to the public in May 2012. We demonstrated some social network features like scale-free and small world effect. Moreover, we investigated the presence of the Homophily principle among the users, and results indicated

that users tend to associate and bond more with others who have similar interests as their own. The findings of these analysis are essential to the understanding of the dynamic change of the structural characteristics of an interest-based social network.

This research was supported by the Government of Abu Dhabi to help fulfill the vision of the late President Sheikh Zayed Bin Sultan Al Nahyan for sustainable development and empowerment of the UAE and humankind.

Acknowledgments

In the name of Allah, the Most Gracious and the Most Merciful.

All praise is for Allah, who has given me the faith, the strength and the courage to complete this work. He has always provided me the guidance throughout my life and made the best choices for me.

I would like to thank my parents and family for supporting and encouraging me to pursue my higher education.

I would like to express my gratitude to my advisor Dr. Iyad Rahwan for his understanding, patience, kindness and support. I am so proud of being a part of his SCAI Lab (Social Computing and Artificial Intelligence Lab) and I thank him for constructing such a collaborative friendly environment among his students. I learned a lot from our meetings and discussions especially from Dr. Iyad's comments to all of the group members. Moreover, I want to thank my research supervisory committee members Dr. Wei Lee Woon and Dr. Zeyar Aung for their constructive feedback and helpful comments on my work.

I want to thank all my colleagues and friends whom I have shared my time with in Masdar. I am grateful for their support and for the memories that we have made in this place. I am indebted to my friends Aamna Al Shamsi, Aamna Al Shehhi and Fatimah Ishowo-Oloko, things would be more difficult without their continuous support and help. I am grateful to Abdulfattah Popoola, Muhammed Aftab and Bijay Neupane for their technical and motivational support and for being

such good friends.

Finally I would like to thank Masdar Institute for this opportunity and every faculty member who have taught me and contributed to my knowledge.

Aisha Al Zbeidi

Masdar Institute, May 26, 2014

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Overview

Online Social Networks (OSN) are the internet-based networks of individuals who communicate and interact in different ways with each other, forming social relationships among them. These networks have been attracting million of users who have adopted these sites in their daily lives activities. As the amount of information and the real-time data uploaded to these systems increasing, it becomes a rich source for researchers from industry and academic institutions to study the human behavior patterns in communities and groups. The structure of these expanding networks is expected to mirror the structure of real-life communications and relationships in society.

In this thesis, we adopted the *Graph Theory* models to analyze the 'Socl' network which is an interest-based social network that is developed by Microsoft FUSE Labs. We considered 'Socl' network as a directed weighted graph where

nodes represent the users or the individuals who participate in the network and the edges represent combination of users' interactions and activities in the network to represent the social tie and relationship among the users.

Either the online or the offline, the social networks are playing a substantial role on shaping the behavior of people in different ways. People adopt many ideas and believes daily that affects their lifestyles, directly from their social connections. Since studying the online social networks in term of global topological structure and behavior patterns between individuals can be comparable to the traditional social networks. The knowledge we acquired from *OSN* could be used in real-life social networks to promote positive sustainable ideas and activities, such as conserving energy, adopting healthy lifestyle or protecting and saving the environment. For example, we can use *centrality measures* to rank and find the most principal and effective actors in a network and then use them to spread the positive idea or effect among the largest component of the network.

1.2 Research Questions

The main contribution of this thesis is to enrich the experimental literature of verifying the theories and models of the social networks proposed by social scientists regarding the social networks features and dynamics over time. However, this thesis investigates the following research questions:

Research Question 1: How does the Socl's topological structure changed over 11 months as it went through 3 different phases?

Research Question 2: How does the differences in the characteristics between the users of every phase contributed to the changes in the Socl network features?

Research Question 3: Does the number of the followers for Socl users correlate with their number of followees? and Is these numbers correlate with the

number of the interests that the users have?

Research Question 4: Is it the different friends who lead the user to get interested in new things. Or is it his interests that shape his friendship selection? And which one of these is the case in our interest-based 'Socl' Network.

1.3 Research Contribution

The main contribution of this thesis is to extend the literature of the experimental studies in the online social networks to verify the proposed theories and models of the social networks in terms of the dynamics and features of the social networks.

This thesis aims to provide a measurement study and analysis on the dynamics of the topological structure and the features of 'Socl' as an interest-based social network. We presented an insight into the dynamic change, where we observed interesting evolutionary patterns on the structure of the 'Socl' network as it went from a private project within Microsoft corporation to a project shared with three universities and invitation only stage before it released to the public. This network grew while going through different phases that are significantly different in terms of size and users' characteristics.

Beside fitting the degree of the users to power law, we noticed the reciprocity behavior in the network considering the strong positive correlation between the in-degree and out-degree corresponding to the number of the followers and the number of followees. Additionally, we investigated the principal of homophily in the network by studying the tend of the users to gain friends with similar interests of their own.

1.4 Thesis Organization

The thesis is organized as follows:

- In Chapter 2 a brief history on the Web Search evolution and transformation over time is presented. A brief overview of some of the literature in the trends of adopting the search activities and the social networks in socializing and collaborating manners for learning purposes are addressed too.

- In Chapter 3 some fundamentals concepts that will be adopted throughout the thesis are addressed.

- In Chapter 4 an overview about the 'Socl', the interest-based social network, is presented briefly. The structure of the raw dataset of 'Socl' is presented, and the preprocessing consideration are discussed too.

- In Chapter 5 the detailed analysis of the structural and topological properties of the network of the subset of active users. Further analysis conducted as well over the 'Socl' network as a network of interests, and we examined the relations between the Interests of the users and their friendship behavior.

- Chapter 6 presents final conclusion and discusses possible implications of our findings on the existed social networks.

CHAPTER 2

Background

This chapter is intended as a brief literature about the Web Search emerging technologies over time. As searching become a substantial activity while using the web, it integrated in many of daily life activities and tasks. Many projects in the literature are considering the collaborative web search for learning and improving productivity purposes among social groups.

This chapter is structured as follow. In section 2.1 a brief history about the web search technologies and how it emerged is presented. Section 2.2 reviews some of the literature in the projects that considered the web search as a collaborative activity among groups instead of a solely activity. Section 2.3 reviews some of the literature in the use of the social networks and the web search for learning purposes.

2.1 History of the Web Search

Since the early development of the web to recent days, the search engines have evolutionized through the time from searching engines relying solely in parsing and looking for the searching keywords in the content of all web pages including the non-related ones, to personalized search engines that can understand and recognize the context of the searching keywords, with considering the users' geo-location and preferences.

When the number of the web servers connecting to the Internet started to increase, it became harder to keep track of the list of the connected web servers and the need of a searching tool has raised. The first basic searching tools in the Internet were using file names and titles to search, such as *Archie* which has developed in 1990 by computer science students at McGill University in Montreal. The Archie searching tool used to maintain a database of anonymous FTP host directories. File Transfer Protocol (FTP) is a standard network protocol that enables any computer to connect to and exchange files with any other computer over the Internet in a client-server architecture. Archie was connecting regularly to the FTP public hosts and downloads the directory listings of the public files of different types such as texts, images, programs, etc. Then, the files were searchable only by their names and titles. Similarly, other searching tools at that point as Veronica and Jughead was developed based on Gopher protocol and used to search file names or titles in the public or local Gopher servers [22].

In 1993, the first web searching engine was developed called Aliweb. It included the location of files on the websites associated with description and keywords submitted by the users. In 1994, the WebCrawler went live, the first full text web searching engine in the World Wide Web that crawl over the whole content of the accessible websites, allowing users to search for any word in any document. WebCrawler is still active search engine owned by InfoSpace, it become

a metasearch engine which combines results from popular search engines, like Google, Yahoo, Bing, Ask.com and other search engines [22].

In 1995, multiple full-text searching engines came out like Alta Vista, Magellan and others that started to rank pages based on the number of occurrences of the words, and a bit about the location of the words in the file with words in the top of the page rank higher than those towards the end of page. The limitation of this approach was the fact that things on the webpages are all under the webpage publisher control. Publisher can continually increase the rank of his webpage and outrank other pages by adding spam lines to increase the number of occurrences of specific words which will misleads the searching results. In contrast, Yahoo was a web portal or directory provided subcategories of interests that maintained by human entries and reviews [22].

In 1998, Google produced a substantial change in the web searching technology through an iterative algorithm called PageRank. Instead of depending solely in the page contents, PageRank suggests a recommendation system that analyzes the relationship between webpages and the links between them. A webpage iteratively ranked based on the number and the Pagerank of the pages that linked back to it. In this approach, the number of other linked webpages will be considered beside their ranks to evaluate the quality of the pages content as a ranking process [22].

In 2007, Universal search has introduced as a new transformation in the web search engines evolution. Instead of the showing list of most relevant webpages and links in the search results, this technology enables a combination of links, images, videos, maps, books and news to show in the searching results. For example, if we searched in the web for 'Abu Dhabi', we will get results combination of city map, images, news, links of official authorities or universities in the city etc. In the PageRank technology, websites publishers were competing based on the text contents of their pages. However, with universal search technology it become

harder since they need to compete not only in the text content of their websites, but they have to consider images, videos, news etc. to be included too [16].

Personalization is another transformation in the web search engines evolution. In 2005, Google launched the personalized search that uses search history to improve the personalized results [13]. In 2008, Google introduced SearchWiki that allows the user to customize his searching experience by re-ranking, deleting, adding and commenting on the search results. Re-ranking enables the user to change the order of the results in the Google's search results page [7]. Searching results are not similar anymore for all users, it has been customized and tailored to best meet the user's needs. Later, Google learned that the people liked the idea of ranking the search results, but didn't like the idea of changing the order of the results in the Google's result page. In 2010, Google introduced the stars feature which replaced the SearchWiki and allowed user to rank the search results based on the personal preferences without affecting the orders of the results in the Google's results page. The starred results are shown later in the relevant searches in a special section at the top of the results page [8]. Country and user location is another way of personalizing results, by finding nearest destinations when search for places or activities in the surrounding area.

In February 2014, Google dominates the market share with 67.5% of search queries conducted in US, followed by Bing with 18.4%, and Yahoo with 10.3% based on the financial analysis exposed by comScore [5].

2.2 Collaborative and Social Search

Web searching has been considered as a single-user activity, which performed in the most commonly used searching interfaces isolated from collaboration and interaction with others. Either a group who is working for a project or homework,

family members or friends who are planning a trip or generally any two or more people who are jointly performing a search activity through the web. All are examples from the daily life that show the need of the collaboration in the web searching activity.

Social search used as a general term that describes when the searchers improve their searching process by taking the advantage of the social expertise networks, social data-mining or collective intelligence [9]. In web searching engines, social search considers the user's social graph by prioritizing the public contents generated or used by any of his social connections when returning results for searches. Evans and Chi [9] defined social search in a wider range, where the social search is an umbrella term that describes the search activities that takes the advantage of the social interactions, while these interactions may be explicit or implicit, co-located or remote, synchronous or asynchronous. They explored the topic of the social search by

As a subset of the social search, collaborative search is a search that is jointly done by participants who collaborate together to meet their information needs [19]. Morris in a recent study surveyed participants with average to expert level of web searching experience [18], and reported that the collaborative web search is a common activity among participants through the searching process and after finding the results. This despite the fact that the existing web browsers and web search engines are not sufficiently supporting the web search collaboration. The need of collaborative learning and working among groups who perform information-gathering tasks in most workplaces led to recent innovations in technologies that support collaborative web searching.

2.3 Searching and Socializing as a Learning behavior

Before the Internet and the Web, we used to rely on human beings to get information or to even guide us in choosing useful references in the literature that will satisfy our information needs. Searching for information is a part of the learning process. Nowadays, the evolution of the web and the search engines and the digital industry shaped the lifestyle of people and changed their learning and educational styles. Huge variety of Information is now accessible to anyone have the access to the Internet, in any time and place. Learners have relied heavily on the search engines to get the information they need from the tremendous amount of information exists on the web. In an exploratory study [12], the authors investigated the possibility of viewing the online searching activity as a learning activity and if it is effective to analyze the web search activities using a learning paradigm. They examined that searchers are depending on the web search as a learning tool, particularly in the cognitive level of learning process called *Applying* where the learner learns how to generalize the knowledge, find commonalities in the different situations and apply it to solve new problems.

CHAPTER 3

Fundamentals

In this chapter, we introduce some of the fundamental concepts in *Graph Theory* and *Statistical properties of Graphs*. In section 3.1 we introduce some graph theory notions, graph properties and centrality measures used throughout the thesis to analyze the So.Cl network. In section 3.2, we explain some theoretical work on the power law property of complex networks.

3.1 Graph Theory

3.1.1 Graph Notion

Conceptually, in graph theory a graph is formed by vertices and edges connecting these vertices. Formally, a graph is pair of sets $G = (V, E)$, where V is the set of vertices and E is the set of edges formed by pair of vertices.

However, graphs could be undirected or directed based on the type of the elements of set of the edges E. Undirected graphs have unordered pairs of elements

in E, where edge (u,v) and (v,u) both represents the same edge formed by vertices u and v. However, Directed graphs have ordered pairs of elements in E, where the edge (u,v) differs than the edge (v,u), despite that both are formed by the same vertices u and v. We use n and m to denote the number of vertices and edges, respectively.

3.1.2 Graph Properties

- **Density**

  The density of the network is the ratio of the number of edges exists in the network to the all possible edges that could be exist if all nodes are connected to each other.

$$Density = \frac{Actual\ Connections}{All\ Possible\ Connections} \tag{3.1}$$

  The density calculated by the equation 3.1 where the number of actual connections is the number of the existed connections in the graph and the number of the all possible connections in the network is calculated by the equation 3.2.

$$All\ Possible\ Connections = \frac{n \times (n-1)}{2} \tag{3.2}$$

- **Shortest Path and Network Diameter**

  In a weighted directed graph, the shortest path is the path with the minimum sum of edges' weights between any pair of nodes. So, if there is a path between nodes u and v that has a set of vertices, then the shortest path noted by $d(u,v)$ defined by the sum of the weights of the constituent edges in path p as shown in equation 3.3 [6].

$$w(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i). \tag{3.3}$$

The average path length is calculated by taking the average of all of the shortest paths exist in the network. The diameter of a network is the longest shortest path among all the shortest paths in the network.

- **Clustering Coefficient**

  The clustering coefficient shows how the neighbors of a random node in the network tend to link together, or in other words, how close the node and it's neighbors to form a complete graph. This measure calculated for every node by dividing the number of triangles containing that node by the number of all possible edges between its neighbors. The clustering coefficient for a single node called 'local clustering coefficient', while the clustering coefficient for a the whole network called 'global clustering coefficient'. The global clustering coefficient calculated by taking the average of the clustering coefficients of all nodes that have at least a degree of 2 [15].

3.1.3 Centrality Measures

In network analysis, the centrality measures represent the relative importance of the nodes within the graph. The centrality information of the nodes in the network reveals important notes on the structure of the network and the data flow patterns in the network.

- **Degree Centrality**

  The degree of the node is the number of direct relationships that the node has[20]. In our example network in 3.1, node A has the highest degree centrality with degree equals 4. Node H has the lowest degree centrality

with degree equals 1.

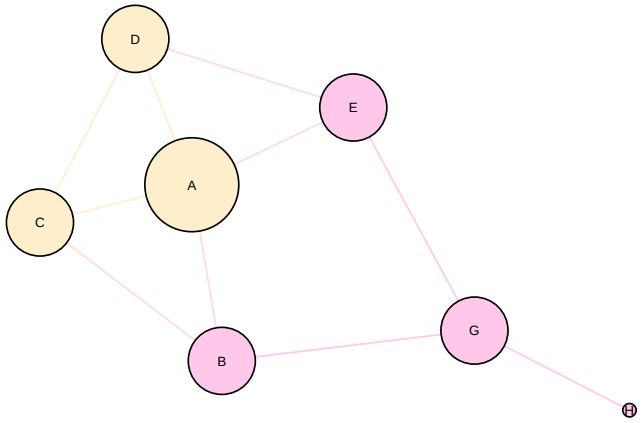

Figure 3.1: An example network diagram where the nodes size represents the node's Degree

- **Betweeness Centrality**

  The betweeness centrality of a node measure how often this node is act as a bridge in the shortest paths between other nodes. In social network this can indicate the nodes that has control over the flow in the network [20]. The betweeness of any node i in a directed graph is calculated by the equation 3.4 where $g_{jk}(i)$ is the number of paths between nodes j and k that passes through node i, while g_{jk} is the total number of paths between j and k.

$$C_B(i) = \sum_{j,k} \frac{g_{jk}(i)}{g_{jk}} \tag{3.4}$$

In our example network in 3.2, we see that node G has the highest betweeness since it is the only node bridging betweeen the connected nodes A, B, C and D with node H. Then, the nodes B and E connects the group of A, C and D to the group of G and H, which gave the nodes B and E higher betweeness values compared to A,C, D and H. The betweeness centrality in *Gephi* calculated based on an improved running-time algorithm provided by Brandes [3].

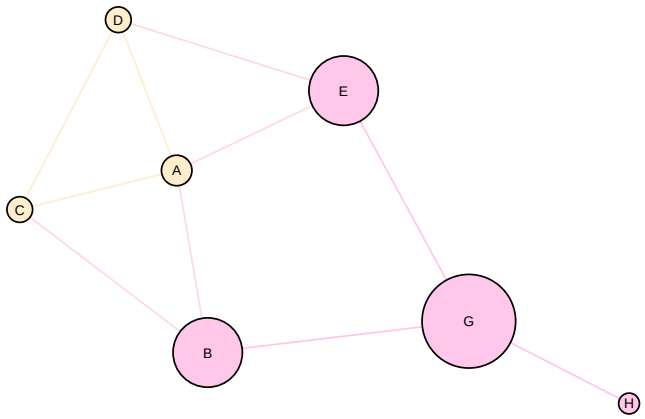

Figure 3.2: An example network diagram where the nodes size represents the node's Betweeness

- **Closeness Centrality** The closeness centrality calculated by inversing the sum of the distances between a node to all other nodes in the network. This metric used to measure how long it will take to spread information from some node to all other nodes sequentially [20]. The closeness of any node i is calculated by the equation 3.5 where we take the inverse of the sum of the all distances from node i to all other nodes in the network.

$$C_c(i) = [\sum_{j=1}^{N} d(i,j)]^{-1} \tag{3.5}$$

In our example network in 3.3 the nodes B and E has the highest closeness followed by A and G. H has the lowest closeness compared to others. The closeness centrality in *Gephi* calculated based on an improved running-time algorithm provided by Brandes [3].

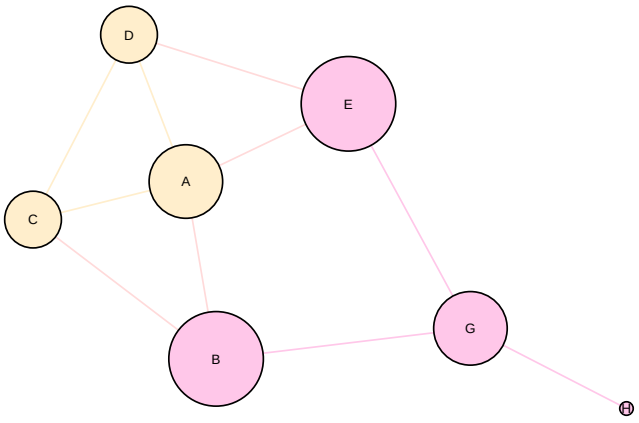

Figure 3.3: An example network diagram where the nodes size represents the node's Closeness

- **Eignvector Centrality**

The eignvector value corresponds to how influential is the node in the network. It ranks the nodes based on the connections of the node. The connections of the nodes to high-scoring nodes contributes more to the rank of the node more than the connections to low-scoring node [20].

In our example network in 3.4, node A has the highest eignvector centrality with eign value equals 1. Nodes B and E follows A with a bit lower eignvalue of 0.96. Nodes C and D has even lower eignvalues, while H has the lowest.

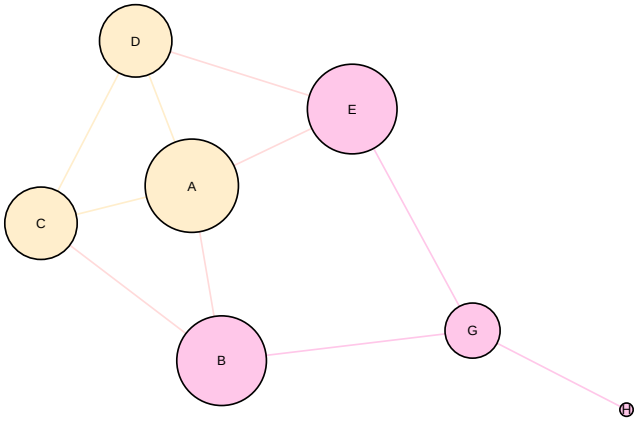

Figure 3.4: An example network diagram where the nodes size represents the node's Eignvalue

3.2 Statistical properties of Graphs

In this section, we will introduce the power law property that are usually detected in several complex networks. In chapter 5 we will examine and analyze the existence of this property in the So.Cl network.

3.2.1 Power law Networks

Across the years the power law has acquired attention due to its productive mathematical properties and the positive physical consequences that accompany it. The populations of cities, the intensities of earthquakes, the sizes of power outages, and examples of power law exists in many of the natural and human made phenomena. In complex networks theory, many researches showed that several real-world networks such as the web, social networks and neural networks are power law networks[4].

For any discrete or continuous variable x has a power law distribution, if the probability density $p(x)$ is proportional to $x^{-\alpha}$. The distribution diverges at zero, so the power law holds for large x, specifically with values of x above a lower bound called $x_{min} > 0$, considering the value of the scaling parameter $\alpha > 1$. Typically, the value of the α between 2 and 3. In practice, the distribution usually doesn't follow a power law for all the values of x, but it does for values of x that is above the x_{min}, in these cases it said that the tail of the distribution follows a power law.

The procedure of assessing the power law distribution that will be followed in the next sections, is provided in Clauset et. al.[4] and summarized as the following:

- Using Maximum Likelihood Method, x_{min} and α values will be estimated for our data that we want to examine its distribution.
- The goodness of fit between the data and the power law will be calculated and the decision to accept or reject will be taken based on the resulted $p-$

value of the goodness of fit.

- Using likelihood ratio test, the power law will be compared to alternative hypothesis. The likelihood ratio value will be used to determine which hypothesis will be preferred over the other, the power law or the alternative.

CHAPTER 4

Socl Network and Dataset

In this chapter, we will present an overview about the Socl network and dataset. In section 4.1, we will introduce the Socl network and how it started as a project in Microsoft Research Labs. In section 4.2, we will discuss the structure of the raw dataset of the 'Socl' network. In section 4.3, we will explain some initial pre-processing steps performed in the dataset prior to the analysis phase which is presented in the next chapter 5.

4.1 Overview about Socl Network

In chapter 4.1, we discussed how the web search engines emerged followed by the social networks. Despite the fact that the most popular social network, Facebook, is a user-based network that centered around people sharing and interacting with their general friends mostly people from their real world, there are many of the social networks experiences which are tailored based on the interests of the users

instead. Examples like Google+, Pinterest and Socl, are interest-based social networks centered around the interests and people are sharing and interacting with people with similar interests.

Socl (http://www.so.cl/), pronounced social, is an experimental interest-based network developed by Future Social Experience (FUSE) Labs division in Microsoft Research and powered by Bing, the search engine. Started as a niche network that targets the students for education and learning purposes. Simply, it is a combination of web browsing, searching service and social networking, where users can express their ideas through visually rich posts that are easy to create, comment on, and share. The main purpose of Socl when started was to explore the possibilities of the collaborative search for learning purposes and to help students and academics who are using social media to enhance their learning experiences by sharing interesting web content as students do when they are working together.

The project started as an internal project within Microsoft and was available only to Microsoft staff members and students from the Information and Design colleges at the University of Washington, Syracuse University and New York University. The invitations started by the mid of November 2011, however; by the mid of December 2011, the project first went public in a limited testing form when the registration was available for new members by invitations only until May 2012. By the end of May 2012, the website opened up to the public and registration become available to anyone using Facebook or Windows Live accounts.

New users can sign up for a new account in Socl using either Facebook or Windows Live accounts. The website will ask at the first sign in for permission to access the user's contacts list and personal information, it will search for friends who have accounts in Socl.

As a starting point, the user have the option to select his interests from the common interest list shared with everybody or from the search tab as shown in

figure 4.1. The user can search for more interests created by other users in the 'Search Tab' or create his own searches and posts by clicking the 'Create Button'.

Figure 4.1: The interface of the Socl website: (1)'Create New Post' button. (2) 'Search Tab' to search for interests by keywords. (3) Suggested List of the most popular interests in the Network.

The user can start away creating posts of different types showed in figure 4.2 in the first box. The second box in the same figure is where the user can choose between three options. The first option is to search for the content on the web using Bing, the second option is to paste the hyperlink of the content to be shared and the third option is to upload the content from the local directory or device. The third box in figure 4.2 is where the user type the topic or the search keyword. The forth box enables the user to specify the type of the results or contents that he want to get such as images, links, videos, or web. The fifth box shows the results of the search based on the type of the contents the user specified. The sixth box is where the user specify the title of the post. The seventh box is the body of the post where the user drags and drops the results and assembles the collage in this case. The eighth box is where the user can add a caption to express the post. The ninth box shows a check-box where the user can choose to share the post on his Facebook wall. The tenth box is the final step to 'Post' or 'Cancel' the post.

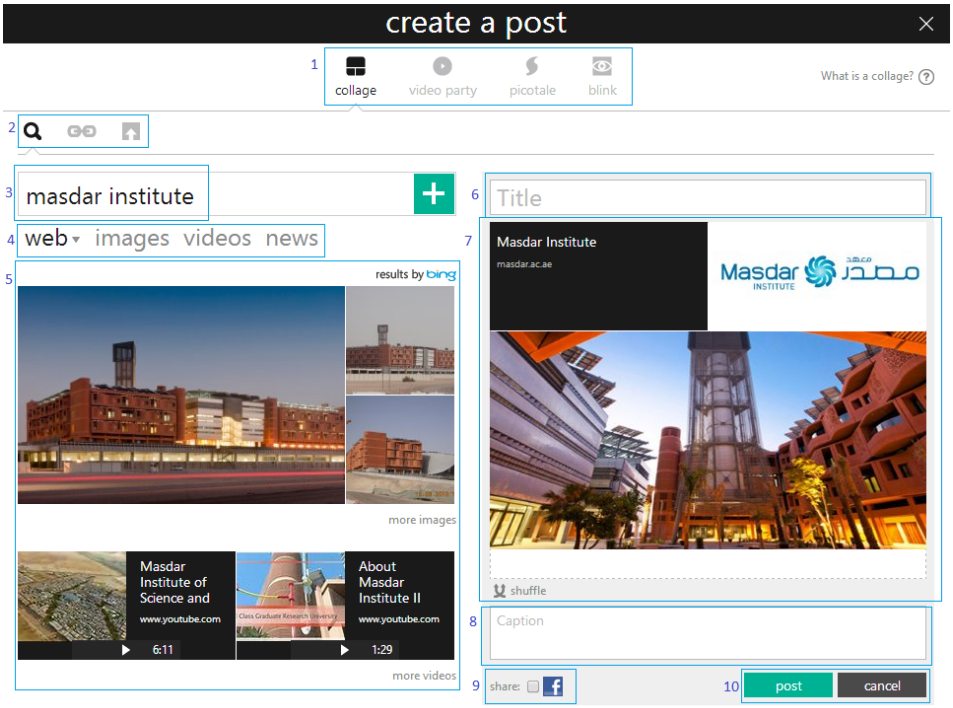

Figure 4.2: The interface of the Create Post in the Socl website: (1) Select the type of the post. (2) Select from three options: search the content on the web, paste a hyperlink of the content, or upload the contents from local directory. (3) Insert topic or search keyword. (4) Choose the type of the results. (5) List of the results from the web. (6) Type the title of the post. (7) Drag from results and drop here to create a collage. (8) Add a caption to express the post. (9) Share on Facebook instantly as it posts in Socl (10) Choose to 'Post' or to 'Cancel' the post.

The posts types are described as follows:

- Collage: is a combination of images, links and videos found on the web or uploaded to 'Socl', created instantly using an easy-to-use drag and drop interface.

- **Video Party:** is a playlist of videos found on the web from sites like Bing, Youtube and Vimeo about a specific topic or interest. The user created the video party and then invite his friends to join with the feature of 'Chatting while Watching', where the groups can watch and discuss the videos in the

real time. This feature is interesting for the learning purposes because it allows a group of people to collaborate and share ideas and thoughts instantly.

- Picotale: is a digital creation of an image overlaid with text. The user enters any message that express his mood or thoughts, and hits *Go*. The website will delivers pictures that reflects or matches the message, the user keeps clicking until he finds the best text/image combination and then post it.

- Blink: is a short interesting and creative dynamic media created using BLINK or BLINK Cliplets apps, available currently for Windows 8 and Windows Phone 8. The user can create and share these clips on Socl or other social networks as well.

The user can view posts in two ways, he can either view all posts of everyone in 'Socl' network, or as he follows more people and interests, he can view his own separate feed page tailored based on his interests and friends feeds. The user then can like, comment or share any post. In figure 4.3 is an example of a 'Socl' post, where the first box shows the name of the user how created the post. The second post shows the type of the post. The third box shows the title of the post and the forth shows the caption of the post. The fifth set of boxes show the 'Like' button with the number of the likes on that post, the 'Comment' button, the 'Collect button' where the user can add the post to his collections and the 'Share' button with a drop-down list of social networks that the user can share on them beside the option to share the posts through emails. The sixth box shows the related 'Tags' of the post. The seventh box shows the comments of the post. The eighth box shows where the user can add a comment. the ninth box is where the user can riff on the post, then the riff will be showed in the comments. Riff is a feature that allows users to collaborate on a post by commenting their further search results in that post. For example, if a user was inspired by another user's post, he can do further

search and add it as a link in the comments of the existing post. The link will refer to the new added post with a small tag in top leads the original post says 'riff inspired by *The Original Post*'. This feature enables groups to collaborate on their searches and results. The tenth box shows the 'Like' buttons for every comment in the post.

Figure 4.3: The interface of the Posts in the Socl website: (1)The name of the user who created the post. (2) The type of the post. (3) The title of the post. (4) The caption of the post. (5) The 'Like' button with a number of likes on it, the 'Comment' button, the 'Collect button' where the user can add the post to his collections and the 'Share' button with a drop-down list of social networks that user can share on them beside the option to share the posts through emails. (6) The related 'Tags' of the post. (7) The comments of the post. (8) The user can add a comment on post. (9) The user can riff on the post, by manipulating the search and re-post as a riff added to the comments. (10) The 'Like' buttons for the comments.

4.2 The Structure of the 'Socl' Dataset

The dataset for Socl is provided by Microsoft Research for researchers at (http://fuse.microsoft.com/research/srd/). The dataset has 3 tables, Users Table, Actions Table and Posts Table, which were loaded to local mysql database for further analysis. Here is a brief about each of them.

1. **Users Table:** contains about three hundred thousand users, registered between September 2011 and November 2012. It has data about when did they join, if they sign in through facebook or windows live, if they joined by invitation from another user or not, finally if the user is a member of the fuse research lab in Microsoft or not.

   By November 2011, according to the dataset, the users were able to register by invitations until May 2012 where the website turned from the invite-only mode to be available to anyone to register and the invitations are no longer applicable.

2. **Actions Table:** contains approximate of 4 million records of about 16 different types of actions such as comments and likes on posts, all are defined in the documentation of the dataset. These actions were taken place between January 2012 and November 2012. This table has information about the timestamp of the action, the source user who did the action, the target user who received the action, the type of the action 'like post, follow user, etc', and the content of the action if available. For example, if the action is following new interest, the content has the name or the title of the interest.

3. **Post Table:** has data about the posts specifically, like timestamp of when it was created, and a timestamp of the last update in the post including any item added to the post, the user who created the post, the user who updates

or adds an item to a post, the type of the item if it is ‘comment, photo, tag, like, video etc’, content of the post and the query content if available.

Figure 4.4 shows the Entity Relationship Diagram of the Socl Dataset.

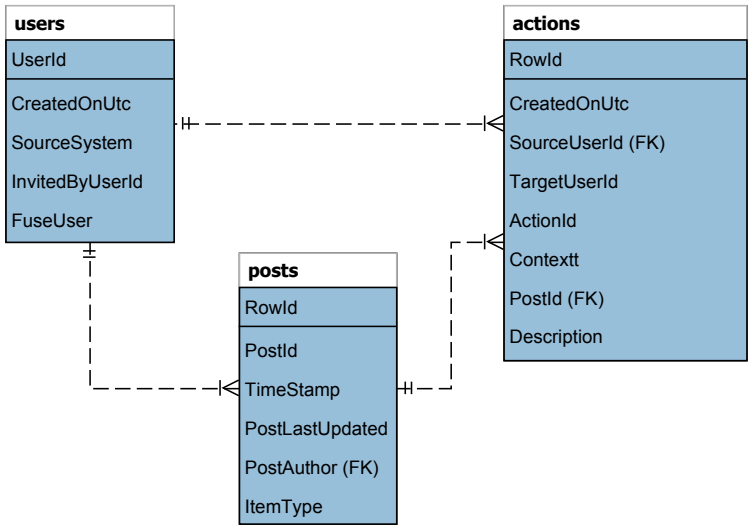

Figure 4.4: The ER Diagram of the Socl Dataset

4.3 Initial Data Preparation in the Socl Dataset

To prepare the dataset to the analysis part, we took in consideration some preprocessing steps over the data. When we first wanted to build the Users Network we faced a problem that we have two conflicted input about when does the users joined the network. The *Actions* table has an action called *Join System* which is a timestamp of when the users first joined the network and it is supposed to reflect the same timestamp in the *Users* table. Considering the review of the ‘Socl’ network project and its expanding over time, ‘Socl’ started as a private project inside Microsoft FUSE Labs since the mid of 2011, to a limited form available to three universities by the end of 2011, and finally available to the public by mid of 2012.

Therefore, we preferred to rely on the *Users* Table which is more reflecting to these facts that we already know about the 'Socl' Network.

The *Users* table in the dataset has the records for users joined the network between the end of September 2011 to the middle of November 2012, however; the *Actions* table started with actions took place between the end of January 2012 to the middle of November 2012. The two tables have a gap of about 4 months, where we have joining users with no actions recorded. To adjust this gap, we started building the Users Network from January 2012 to November 2012, where for each month the network was updated by adding the joining users extracted from the *Users* Table and forming the connections between them using the actions that took place in that month extracted from the *Actions* table.

As the records in the dataset have associated timestamps, we used it to solve the problem of redundant actions that we can accept only once in our model such as *follow another user* action where we consider only the first occurrence of the action for a distinct source and target users. For example, if user A followed user B, we will have a record shows the action's source, target and timestamp in the *Actions* table. However, if user A for some reason un-follow user B and then re-follow again, we will consider the first action only and ignore the another one.

4.4 The Structure of the Users Network

4.4.1 Modeling the General Structure for the Socl Network

It is important step to choose a representative model for the network, even though it may look simple, but it is not so trivial. The friendship concept in the 'Socl' network is *unilateral*, where it is not necessarily for users to request or confirm the relationships among them. For example, if user A follows user B, that doesn't imply either user B will confirm this connection or will be following user A as

well.

The connections between users are *directional*, where there is an individual who initiate the action and identified as the source, and there is an individual who receive the action and identified as the target. In Socl, it is not required for user A to follow user B to be able to comment, or like any of user's B posts, and these communications doesn't necessarily imply that one of them has followed the other. Therefore, to model a representative network of connections between the users, we should not consider the *'Follow another User'* action solely and ignore the amount of the *'Comment and Like'* actions between the users.

Therefore, I considered the Users Table to build the Users Network by extracting data about users, when they *Join the Network*. To measure the amount of interactions between the users, I considered mainly the *Actions* table to extract data about the actions *follow another user*, *Add a comment on a post* and *like a post* when each occurrence for any of these actions strengthen the connection between the source and target users. In the case of users who were invited to join the network by another user, I used the *Invited by* column from the *Users* table, where this invitation adds to the edge weight between the invited user as a source to the user who sent the invitation as a target.

Our model for the Network represented as a graph $G = (V, E)$, where nodes V represents the users and edges E represent the connection between the nodes and the weight of the edges represent the amount of the communication between them. The graph is considered directed weighted graph, where the edge direction identified by the source and target users of the actions, and the edge weight identified mainly by the frequency of these three actions *follow another user*, *Add a comment on a post* and *like a post*, where every occurrence of action assigned a weight of 1 adds up to the edge weight. As we mentioned earlier, in case of the joined user was invited by another, this will be counted once in the edge formed from the

invited user to the user who sent the invitation. Finally, we considered the self-loops not allowed in the model, ignoring all the records that has similar source and target User ID in the *Actions* table, for example if a user commented or liked his own post. Figure 4.5 shows an example scenario of the weighted directed edges representation for the communications between users A and B the users network.

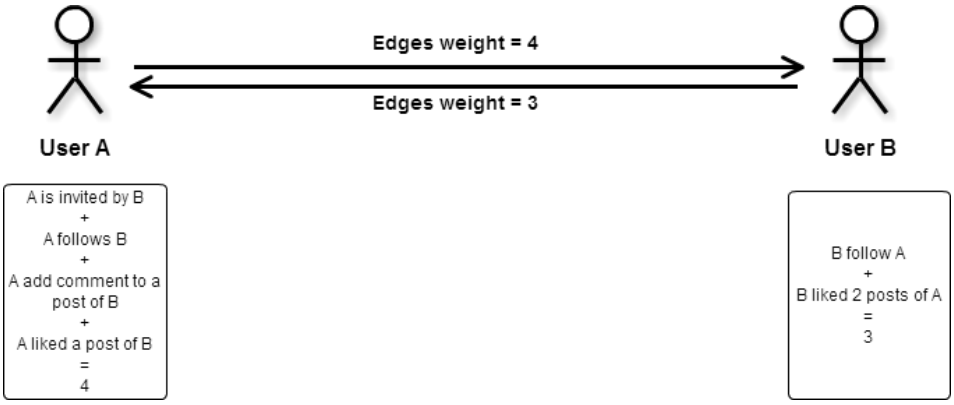

Figure 4.5: Example scenario of a weighted directed graph of two nodes A and B where the edges weight calculated mainly by summing up the number of occurrence of any of these three actions *follow another user*, *Add a comment on a post* and *like a post*. In case of a user was invited by another, this will adds a weigh of 1 to the weight of the directed edge from the invited user as Source node

4.4.2 Specified Model for the Active Users in Socl Network

The 'Socl' network is a special interest-based social network intended for learning and collaboration purposes, changed over the time between September 2011 and November 2012 due to the fact that it had moved through different phases over this period of time. It is not trivial to capture the change of the users behavior patterns in the network while the type of the users themselves was changing in each phase. The first phase before December 2011, the project was private and the users were Microsoft staff including the FUSE lab members. In the second phase, the project was opened to students of three different universities, besides the invited users who are most probably students or students' like-minded fellows. The third phase, the

network opened up to the public from everywhere to join, which represents mostly very different type of users and maybe different behavior pattern.

To capture the effective part in the network and analyze it, we choose to select the active users defined by the general model of the 'Socl' network with additional filtering procedure to exclude the non-active users which are not considered effective. This sub-network of the active users is the network to be considered in the analysis in cahpter 5. The filters to be applied to generate the *Active Users* network specified as the following:

- Giant Component: is the most important part of the directed graph and it is the first filter applied to our network. The giant strongly connected component is where every pair of vertices is connected in both directions,for example, from one of the vertices, one can approach the other by moving either along or against the edge directions.

- Node Degree: is the sum of the in-degree and out-degree as we have a directed network. The threshold for nodes to be considered as active nodes is to have at least a degree of 3. A simple scenario to justify this value is if a user A was invited by user B and joined the network, the first intuitive action once user A will join is that both of them will follow each other, which represents a degree of 2 for both users. This user can't be considered active yet until he make additional edge with different user which shows that he is interested to be active in the network.

- Edge Weight: as described in our model is the count or the frequency of the actions that are initiated from a source to target user. In the Active Users network we specify the minimum threshold for the edge weight to be 2. In the same scenario we used to justify the threshold for the node degree, the weight of the edge from B to A is 1 where the only action taken place which

is *follow another users*. This can't be considered as an active connection or edge that involves collaborative communication. Therefore, as the edge weight go further that indicates that the user become more active.

CHAPTER 5

Analysis

In this chapter, we will analyze some of the structural properties of the Active Users network, that are usually detected in several social networks. Moreover, we will study the general structure of the Interests network and explore the homophily among the users based on their interests.

Section 5.1 provides the detailed analysis on the active users network. In 5.1.1 presents a general statistics on the network's size changes through the time between January and November 2012. In 5.1.2 we explore the communities dynamics in the Socl network. In 5.1.3 we examine the degree distribution dynamics, first we tested the power law fit to the degree distribution and then we measured the correlation between the in-degree and out-degree of the users. In 5.1.4 we study the dynamics and changes on the average clustering coefficient and the diameter of the network.

Section 5.2.1 provide the general structure of the Interests network. Section 5.2.3 tests the correlation between the user's friendship and his number of interests.

In section 5.3 of this chapter, we explore the homophily existance in the 'Socl' network.

5.1 Analysis of the Users Network Structure

To compute and visualize the growth and changes of the users network structure over time, we used *Gephi* the open-source networks visualization software [1]. It is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. It is a free tool which runs on Windows, Linux and Mac OS X and it has built-in functions to calculate network properties and metrics such as clustering coefficient, centrality and community detection. We used these built-in functions in *Gephi* to compute the networks properties for our analysis.

Furthermore, we used *python* combined with *mysql* queries to produce refined files to be used in the further statistical analysis. We used *RStudio*, the IDE for *R* the statistical language, to conduct some graphics and statistical computing analysis on the Socl dataset.

5.1.1 Network Size Growth

Over the 11 months range of time, we monitored the growth in the network size and the changes on the network metrics. Table 5.1 shows the high-level statistics of the Socl dataset and the growth in the network size represented by the aggregate number of nodes and edges in a monthly basis between January and November 2012.

	All Users Network		Active Users Network	
Range of Time	No. Nodes	No. Edges	No. Nodes	No. Edges
Jan. 2012	17355	36180	241	1242
Feb. 2012	23324	42494	405	1674
Mar. 2012	25878	45688	462	1830
Apr. 2012	27811	48420	805	3248
May. 2012	247590	391673	3397	16625
Jun. 2012	279706	533996	4684	30571
Jul. 2012	290191	581695	5124	35185
Aug. 2012	300090	630996	5549	40232
Sep. 2012	305902	653595	5768	42908
Oct. 2012	310348	670295	5939	43786
Nov. 2012	312211	678340	5970	44196

Table 5.1: The Users Network Expansion through the time between January and November 2012 for the Network of All Users and the Network of Active Users: The Cumulative Number of Registered Users and Approximate Number of Connections among them. The Nodes and Edges in the subset Network of Active Users are for Nodes who considered *Active* and satisfy these two conditions: Node Degree > 2 and Edge Weight > 1. The average percentage of the proportion of the number of nodes in the subset network to the whole network nodes is about 2%. The average percentage of the proportion of the number of edges in the subset network to the whole network edges is about 5%

5.1.2 Communities Dynamics

In this section we will present the growth of the network considering the communities formed in the range of time between January and November 2012. As we discussed earlier how Socl network started and went through different stages each with different purposes and type of users. In this section, we will provide an overview over these stages or phases of the Socl network project and how moving through these phases affected the way that communities formed.

The community detection algorithm that we used to detect the communities in the Socl network is Louvain community detection algorithm, which also called Louvain modularity method [2]. This algorithm is based on modularity optimization and it outperforms other algorithms in terms of computation time. Moreover,

it's quality of detecting the hierarchical structure of the communities is verified on existing networks and it detects in different resolutions that could be specified in advanced. Comapred to other known community detection algorithms, this algorithm achieved better performance measured by the so-called modularity. The mechanism of this algorithm has two phases, in the first phase all nodes are assigned to different community so we have communities as much as nodes. The next phase is to iterative, for every node i with neighbors j, we check the gain of modularity of the node i if removed from its community and added to the community of j, the community that have the maximum modularity gain is chosen.

The network layout algorithm that was used here is an optimized version of what is called ForceAtlas algorithm. ForceAtlas2 algorithm is a force-directed algorithm with real-time adjustable settings that affect the shape of the network such as scaling, speed, gravity and repulsion [11]. It has a linear-linear model where the attraction and repulsion are proportional to the distance between the nodes and the unique adaptive convergence speed allows most graphs with up to 1 million nodes to converge more efficiently.

The algorithm that we used to evaluate the impact of the nodes in the network is called HITS or Hubs and Authorities algorithm [14]. This algorithm evaluates the importance of a nodes by considering the importance of its neighbors. By iterative process, nodes with neighbors of high authority get higher authority values.

Phase 1: Socl as a private project within Microsoft

In this phase the users are the staff of Microsoft including members of the FUSE labs. Based on the 'Socl' dataset, this phase took place between September 2011 and the end of December 2011. Due to the fact that we have only the *Actions* table records started only from January 2012, we visualized the active users network in January 2012 in figure 5.1. However, we can notice that the network has couple

of communities which is expected to represent the communities or the departments within Microsoft beside some new joining users. Given the IDs of the users who are members in the FUSE lab, we can see most of them forming a community, in yellow color, in the center of the network which has nodes ranked as the highest authorities in the network.

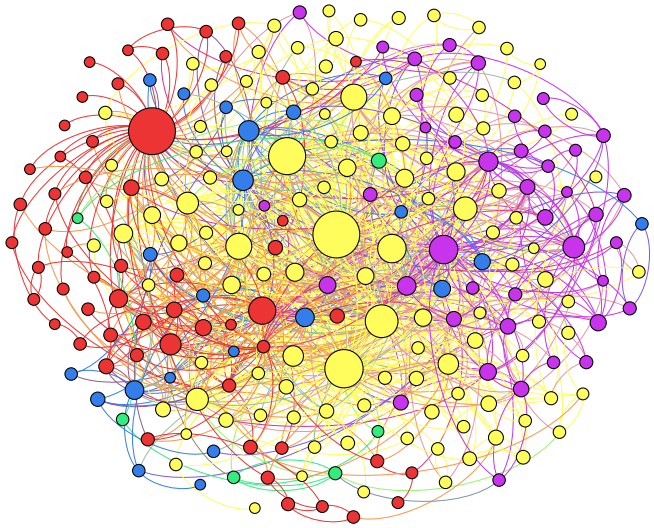

Figure 5.1: The network of active users in January 2012 visualized using Gephi: The colors of the nodes represent the communities, the size of the nodes represents the authority value calculated by Gephi using HITS algorithm. The node's authority value represents how good is the node as a source of information

Phase 2: Socl as a limited testing form

In this phase the new joined users are students of Information and Design colleges from three Universities. This phase took place between the end of December 2011 and the end of May 2012. In February 2012, the students who joined the network formed a new well-connected community that has only few connections to the community of the first phase users. In March 2012, new users formed an additional well-connected community, and similarly in April 2012, new users formed an additional new community too. By May 2012, as shown in 5.2 there were three

communities formed in the second phase, beside the community of the first phase. Due to the fact that in this phase the network was open for students from three different universities, this can explain the three well-defined and well-connected groups formed in this phase. The central node in figure 5.2, is the admin of the system and the only direct connection between the community from the first phase shown in the left and the three new communities to the right side.

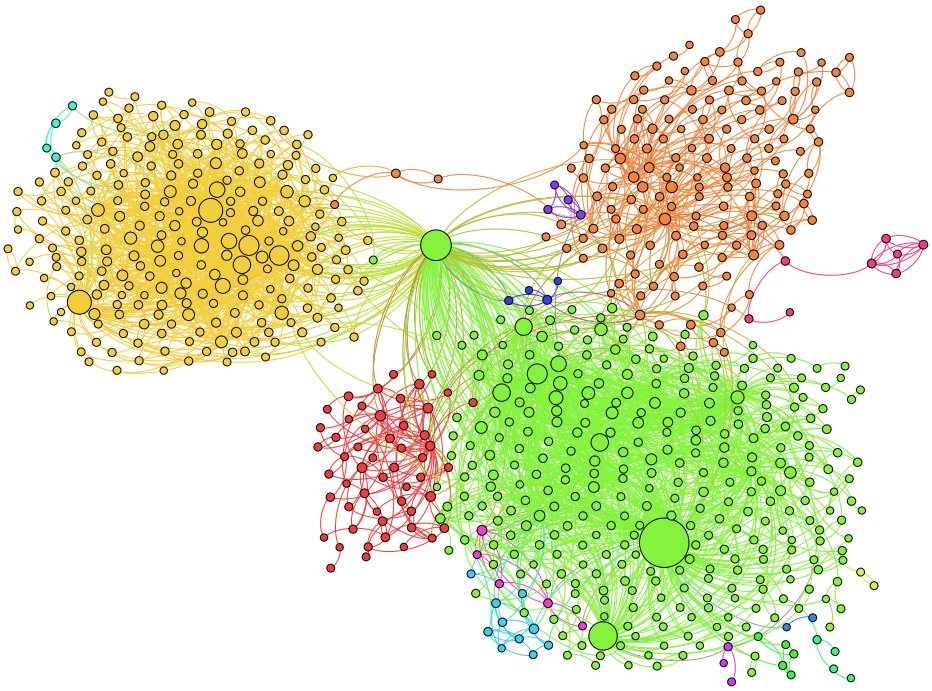

Figure 5.2: The network of active users in April 2012 visualized using Gephi: The colors of the nodes represent the communities, the size of the nodes represents the authority value calculated by Gephi using HITS algorithm. The node's authority value represents how good is the node as a source of information. Users from Phase 1 shown to the left of the central node as one community. The main three communities of Phase 2 shown in Green, Red and Orange

Phase 3: Socl as a public social network

In this phase the users are anyone who has either Facebook or Windows Live accounts to sign up with. This phase took place from the end of May 2012 until

now. Based on our dataset, we will be analyzing this phase in a monthly basis from May to November 2012. Around 20th of May, Socl went public and the size of the network increased tremendously. This phase is significantly differs from the previous two phases, since the network is not as straight forward to visualize given the huge number of joined users who are not necessarily to be users with common characteristics. However, we believe this phase could be more representative than the previous phases to the users diversity and the behavior trends that exist in the popular social networks. Figure 5.3 shows the Socl network communities in the mid of November 2012. The groups of users from the first and second phases become peripheral by the end of the third phase, while the new users were the main influential group clustered in the centre of the network. The reason that made the groups of the first and second phases become less or not active could be the fact that the network was in a testing stages in the first and second phases which makes the users not interested anymore to continue their participation.

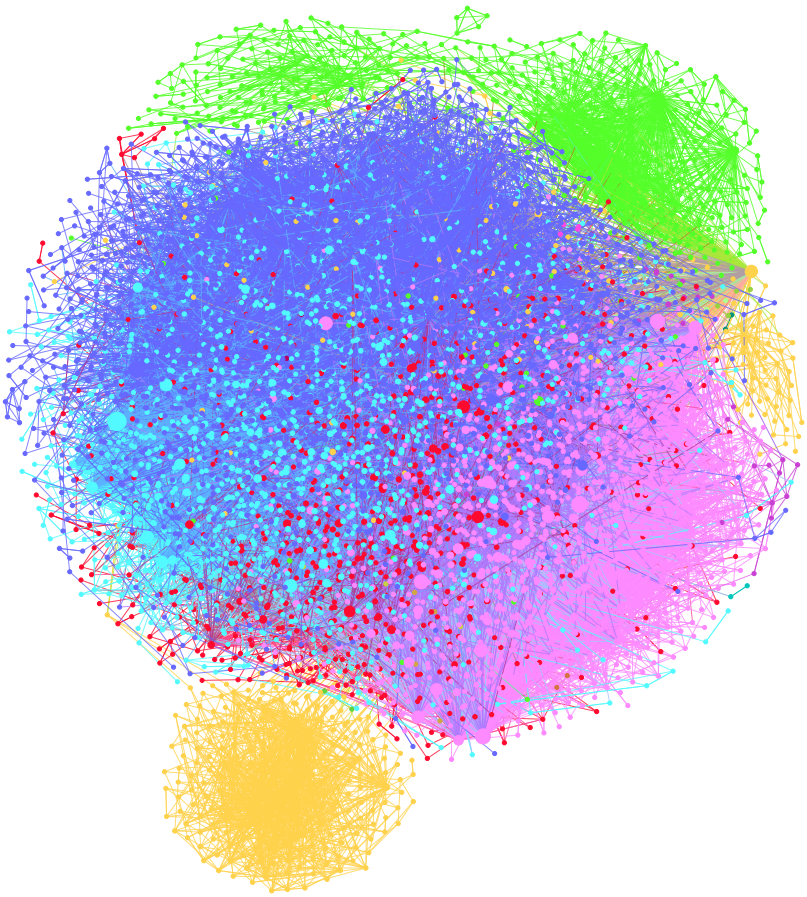

Figure 5.3: The network of active users in November 2012 visualized using Gephi: The colors of the nodes represent the communities, the size of the nodes represents the authority value calculated by Gephi using HITS algorithm. The node's authority value represents how good is the node as a source of information. Users from Phase 1 are colored in yellow 'Lower to the network', while users from phase 2 are colored in green 'Upper to the network'. The central huge network are users from phase 3, clustered in different groups in light blue, dark blue, red, and purple. User with id '1' is connected to the old and the new groups, shown within a small yellow group to the right of the new bigger network

5.1.3 Degree Distribution Dynamics

In this section, we analyze the degree distribution of the nodes throughout the three different phases. In this analysis we considered only the follow action in forming the connections between users and discarded the affect of the unfollow action. In the analysis, the in-degree and the out-degree distributions represents the number of followers and the number of followees distributions respectively.

Fitting the Degree Distributions to Power-Law

In this section, we will be testing for the active users network whether the distribution of the In-degree, Out-degree and the Degree as a sum of both, follow a power law distribution. Power law distributions are usually used to model data whose frequency of an event varies as a power of some attribute of that event. In our case, we will see if the frequency of the in-degree and the out-degree of users vary as a power of the in-degree and the out-degree itself. In other words, we will see if there are few in-degree and out-degree values that are very common and if there are many in-degree and out-degree values that are not as common.

We used the data of the subset of the active users network which has nodes with at least a degree of 3, and edges with at least weight of 2. We used the *poweRlaw* [10] package in *R* language, to fit the Number of followers and followees distribution to the power-law distribution. This package used the algorithms presented by Caluset et. al[4]. As a reminder of the power law analysis procedure mentioned in 3.2.1 that we are going to follow in this section, we will follow the following steps:

- **STEP1:Estimating the Power Law Parameters**: Using Maximum Likelihood Method, x_{min} and α values will be estimated for our data that we want to examine its distribution.

- **STEP2: Goodness of Fit tests**: The goodness of fit between the data and

the power law will be calculated and the decision to accept or reject will be taken based on the resulted $p-value$ of the goodness of fit.

- **STEP3: Alternative Distribution**: Using likelihood ratio test, the power law will be compared to alternative hypothesis. The likelihood ratio value will be used to determine which hypothesis will be preferred over the other, in our case, whether the power law or the alternative.

STEP1: Estimating the Power Law Parameters

First step is to estimate the two important parameters x_{min} and α. The approach for estimating the best value of lower bound of the power law behavior x_{min} is to estimate the value that makes the both probability distributions of the measured data and the best-fit power law as similar as possible. Estimating x_{min} is the tricky part in this step, because if the estimated x_{min} was lower than the true x_{min}, the distribution will deviate at great rate from the power law model and there will be a fundamental difference between the data and the model. However, if x_{min} was chosen to be higher than the true x_{min}, the sample size will be reduced significantly which will affect the distribution and make it less matching to the model. The procedure in *poweRlaw* [10] package involves the use of Kolmogorov-Smirnoff(KS) statistic, that measures the maximum distance between the Cumulative Density Function (CDF) of the data and the fitted model, in which it finds the x_{min} that minimizes the value of the KS statistic. After finding the value of x_{min}, we use the Maximum Likelihood Estimation MLE to compute a plausible value for α.

Table 5.2 shows the estimation parameters of fitting the In-degree, Out-degree and Degree distribution to a power law model.

From table 5.2, we can see a pattern in the power law parameters that change as the network moves from phase to another. Typically, the value of α is in the range between 2 and 3, where the the mean exists but the variance and higher-

	Indegree			Outegree			Degree		
Active Users in	x_{min}	α	KS	x_{min}	α	KS	x_{min}	α	KS
Jan. 2012	2	1.98	0.0555	3	2.01	0.0749	3	1.9	0.0554
Feb. 2012	6	3.16	0.0271	12	2.48	0.0420	7	2.96	0.0139
Mar. 2012	6	3.16	0.0280	12	2.46	0.0459	7	2.96	0.0139
Apr. 2012	6	3.16	0.0315	8	2.46	0.0417	7	2.89	0.0166
May. 2012	7	2.96	0.0111	5	3.03	0.0225	9	3.05	0.0173
Jun. 2012	6	2.76	0.0111	5	2.97	0.0221	23	2.59	0.0134
Jul. 2012	6	2.72	0.0115	31	2.41	0.0225	34	2.48	0.0176
Aug. 2012	6	2.70	0.0124	29	2.39	0.0234	35	2.43	0.0162
Sep. 2012	6	2.69	0.0113	31	2.34	0.0196	37	2.4	0.0159
Oct. 2012	6	2.68	0.0123	31	2.34	0.0213	37	2.4	0.0154
Nov. 2012	6	2.68	0.0108	29	2.35	0.0212	39	2.39	0.0164

Table 5.2: The Parameters Estimation of Fitting Power law to the In-degree, Out-degree and Degree Distributions between January and November 2012: The data fitted to power law above the lower bound x_{min} that is the optimal value that minimizes the value of the Kolmogorov-Smirnov test value KS which is the maximum distance between the CDF of our data and the CDF of the theoretical power law model, where the scaling parameter of the power law is α estimated using MLE. The results calculated using the R implementation of a power-law distribution fitter found in http://tuvalu.santafe.edu/~aaronc/powerlaws/plfit.r

order moments are infinite [23]. In our analysis the α varies between 1.98 and 3.16 for the in-degree, and between 2.01 and 3.03 for the out-degree and between 1.9 and 3.05 for the degree which are still considered reasonable values for α.

In general, the results showed a trend of degree distribution changes between the three phases. Therefore, to examine the goodness of this power law fit, we will consider a month from each phase to represent the degree distribution for that phase. The results in next step will identify whether the power law fit could be a plausible hypothesis for these distributions or not in every phase.

STEP2: Goodness of Fit tests

Second step is to test how well our power law distribution fits our data, by using the Kolmogorov Smirnov (KS) test. Using our estimated power law parameters from the first step, large number of power law distribution synthetic data sets can be

generated and compared to our data distribution to see if our data and the generated data come from the same distribution. The method used is to compare each of the generated data sets to its own power law model by performing a *KS* test.

Then the p-value decided by counting the fraction of the time that the resulting test value is larger than the empirical data. We test with 0.1 as the significance level, which means that only 10% or less of our *KS* tests will reject the null hypothesis that the power law is plausible. Therefore, if the p-value is greater than 0.1 we fail to reject the null hypothesis and we conclude that both data sets come from the same distribution. We conclude that the power law with the given parameters is a plausible hypothesis with significance level of 0.1. In table 5.3, we provide the goodness of fit results for January, April and June, each month from different phase.

	Indegree		Outdegree		Degree	
Range of Time	$p-value$	gof	$p-value$	gof	$p-value$	gof
Jan. 2012	0.536	0.04014	0.088	0.06924	0.0792	0.05496
Apr. 2012	0.002	0.03410	0.066	0.04407	0.1576	0.01653
Jun. 2012	0	0.01411	0.026	0.02259	0.236	0.014409

Table 5.3: The Goodness of Fit tests for In-degree, Out-degree and the Degree Distributions. The p-value and goodness of fit values are found using the results of 2500 *KS* tests performed on each of the samples of the synthetic data sets which are sampled from a true power-law distribution. If the resulting p-value is greater than the significance level of 0.1 the power law is a plausible hypothesis for the data.

Based on the results in 5.3, for the in-degree distribution, the power law is a plausible hypothesis only in January where the p-value is greater than 0.1. However, for months April and June, the p-value is less than 0.1 so we reject the null hypothesis and we conclude that the in-degree distribution and the power law model doesn't come from the same distribution in both April and June networks. Furthermore, the p-value is less than 0.1 for the out-degree distribution in January, April and June, so we reject the null hypothesis and we conclude that the power

law model and the out-degree distribution doesn't come from the same distribution. Finally, for the degree distribution which is the sum of the in-degree and the out-degree combined, the p-value is less than 0.1 in January, where the power law fit is not a plausible hypothesis. However, in April and June, the p-value is greater than 0.1 so we fail to reject the null hypothesis and we conclude that the power law and the degree distribution come from the same distribution in the second and the third phase.

STEP3: Alternative Hypothesis

In previous step, we already tested whether the in-degree, out-degree and degree distributions are plausibly drawn from a power-law distribution, and we found out that the power law fit can't be ruled out for the in-degree distribution in the first phase and for the degree distribution in the second and third phase.

However, if our distribution is well fit by a power law, that doesn't really imply that the power law is the best fit, we can still find better fits using another distribution. We can test this possibility of finding alternative better fit by using a goodness-of-fit test again by calculating a p-value for a fit of the competing distribution and compare it to the p-value for the power law.

However, we will use the direct comparison method to compare two models to find out the better fit among them. The method uses the likelihood ratio test which calculates the likelihood of the data for the two competing distributions and the better fit is the distribution with the higher likelihood value. Using the *poweRlaw* package[10], we directly compared the power law against the log-normal distributions using the ratio of the two likelihood or equivalently the logarithm of the ratio (*LR*). The sign of the resulted *LR* value can indicate which distribution is better, but with considering that the value of *LR* should be a sufficient value that is not close to zero so its value couldn't be affected by the statistical fluctuation. The negative value of the log likelihood ratio *LR* indicates that the alternative distribution is fa-

vored over power-law model. The p-value indicates if we reject the null hypothesis or not, while the sign of the *LR* value indicates which model fits better to the data.

By comparing the power law and the log-normal fits to the in-degree distribution in January 2012, we get the Likelihood ratio *LR* equals to (-2.109) and the $p-value$ equals to (0.0175). The p-value is less than 0.1 which is statistically significant in this case, and it means that the sign of the value of *LR* is unlikely to be a chance result of fluctuations and the sign is reliable indicator to which model is better fit to the data. Considering that the value of *LR* is sufficiently negative that we make sure we avoid the statistical fluctuation in the decision, so we conclude that we reject the null hypothesis and conclude that the log-normal distribution is better fit to the in-degree distribution in January 2012.

In figure 5.4 we see the power law in red fits to the In-degree distribution compared to the log-normal in green as a better fit.

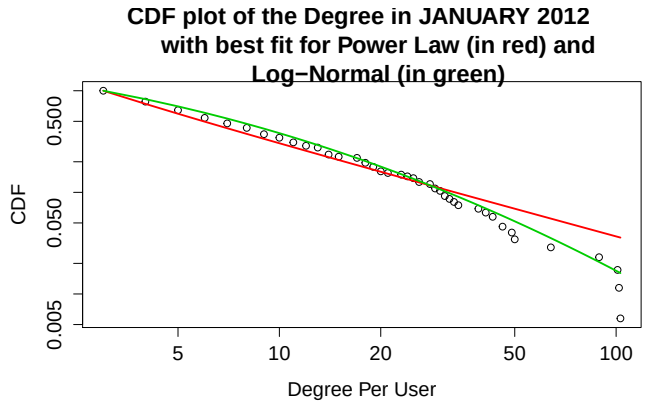

Figure 5.4: The CDF plot of the degree for January 2012, with comparison of both power law and log-normal distributions fits, with log-normal as a better fit

For the degree distribution in April and June 2012, we compare again the power law and the log-normal fits to see which gives better fit. For April, we get the Likelihood ratio *LR* equals to (+0.634) and the $p-value$ equals to (0.737). The p-value is larger than 0.1, however the *LR* value is not sufficient value which means

that we can't rely on as an indicator of which model is better fit. In figure 5.5 we see the degree distribution with the power law as a better fit compared to the log normal.

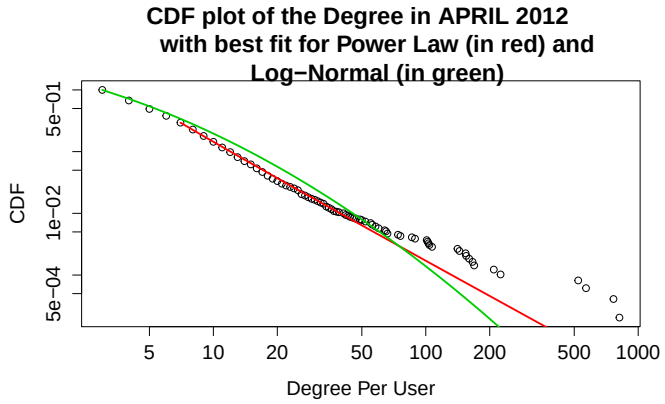

Figure 5.5: The CDF plot of the degree for April 2012, with comparison of both power law and log-normal distributions fits, with power law as a better fit

For June, we get the Likelihood ratio LR equals to (+1.1) and the $p-value$ equals to (0.864). The p-value is greater than 0.1 which is statistically significant, and it means that the sign of the value of LR is unlikely to be a chance result of fluctuations and the sign is reliable indicator to which model is better fit to the data. Considering that the value of LR is sufficiently positive value that we make sure we avoid the statistical fluctuation in the decision, we fail to reject the null hypothesis and we conclude that the power law is favored over the log-normal distribution in June 2012. In figure 5.6 we see the degree distribution with the power law as a better fit compared to the log normal.

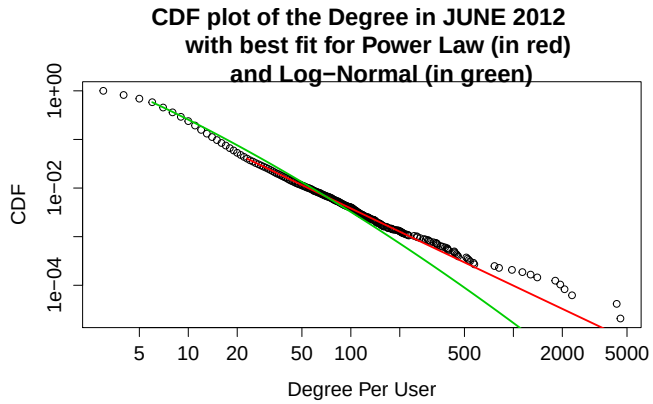

Figure 5.6: The CDF plot of the degree for June 2012, with comparison of both power law and log-normal distributions fits, with power law as a better fit

To measure the inequality in the degree distribution, we used the gini-index. Table 5.4 shows the gini coefficient values from January to November 2012. The range of the values of the gini coefficient is between 0 and 1, where the coefficient of 0 represents the perfect equality, while the coefficient of 1 represents the maximal inequality in the distribution. In our case, the high values of the gini coefficient indicates the high inequality among the degree values of users in the 11 months.

Active Users Network in	Gini Coefficient
Jan. 2012	0.994
Feb. 2012	0.581
Mar. 2012	0.622
Apr. 2012	0.610
May. 2012	0.809
Jun. 2012	0.799
Jul. 2012	0.801
Aug. 2012	0.802
Sep. 2012	0.803
Oct. 2012	0.803
Nov. 2012	0.804

Table 5.4: The Gini-index for the Degree Distribution from January to November 2012

Degree Correlation

In this part, we will explore the correlations between the Number of followers and followees for the active users in the Socl Network. To compute the correlations, we used Spearman method which is a rank correlation method that uses a monotonic function to measure the dependence between the variables without requiring such relationship to be represented linearly. This correlation method is suitable in our case because it is less sensitive to the non-normality distributions. Table 5.5 shows the correlation coefficients for the network from January to November 2012.

Active Users Network in	Correlation Coefficient
Jan. 2012	0.87
Feb. 2012	0.25
Mar. 2012	0.31
Apr. 2012	0.27
May. 2012	0.45
Jun. 2012	0.48
Jul. 2012	0.48
Aug. 2012	0.48
Sep. 2012	0.49
Oct. 2012	0.49
Nov. 2012	0.49

Table 5.5: The Correlation Coefficient between the in-degree and out-degree for the active users network in 'Socl' between January and November 2012. The correlation method used in this calculations is Spearman method [21]

From table 5.5 we notice that the in-degree and the out-degree in the first phase have very strong positive correlation, which suggests that the number of followers of a user is strongly related to his number of followees. Therefore, users with increasing number of followees tend to have increase in their followers as well. In the second phase the correlation coefficient decreases significantly to be weak positive correlation. However, the correlation increased again to represent the strong positive correlation between the number of followers and followees for the active users in the third phase.

5.1.4 Diameter and Clustering Coefficient Dynamics

In this section, we will examine the small-world property of the social networks in the active users network.

As we tested in the previous section the power law fit to the degree distribution and we found that the power law is a good fit as the network increased in the size as in the second and third phases which indicates the existence of the scale-free property in the network. We also tested the inequality of the degree distribution using the gini index and found that we have high inequality in the active users network. That indicates the high centrality in the network as we have few nodes that are connecting to most of the nodes in the network.

In this part, we will discuss the change on the diameter and the network's average coefficient over the months from January to November 2012. Table 5.6 shows the changing of the active users network's diameter and the global average clustering coefficient. The average clustering coefficient of the network ranges between 0.135 and 0.210, with the average overall all months equals to 0.176. Average clustering coefficient in this high range suggests the strong local clustering in the network, which could be explained by the tend of the users to connect to their friends' friends [17].

Active Users Network in	Avg Clustering	Diameter	Average Shortest Path
Jan. 2012	0.23	8	3.03
Feb. 2012	0.208	12	4.147
Mar. 2012	0.208	12	4.237
Apr. 2012	0.210	13	4.532
May. 2012	0.135	13	4.368
Jun. 2012	0.152	13	4.104
Jul. 2012	0.152	13	4.097
Aug. 2012	0.160	13	4.071
Sep. 2012	0.169	13	4.038
Oct. 2012	0.160	13	4.045
Nov. 2012	0.161	13	4.042

Table 5.6: The global average clustering, the diameter and the average shortest path of the active users Network between January and November 2012

Figures 5.7 and 5.8 show the average clustering coefficient plotted as a function of the nodes' degree for the network in January and April 2012 respectively. The clustering coefficient is higher for nodes with low degree, which suggests that users with low degree are tightly clustered.

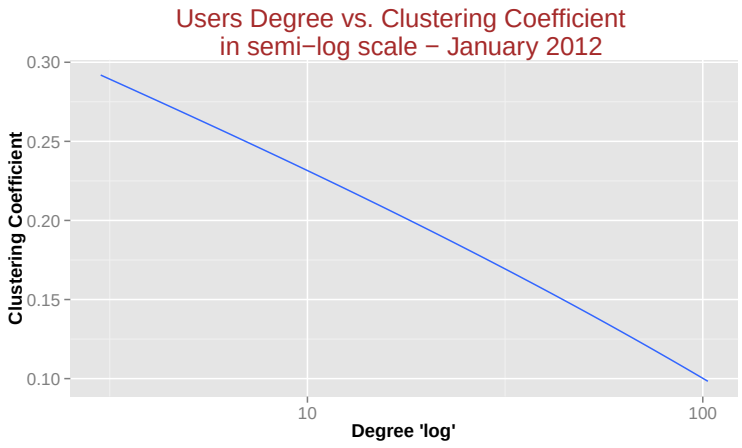

Figure 5.7: Clustering Coefficient of users with different degrees in JANUARY 2012. The users with few friends are tightly clustered while the users with many friends are less clustered. The graph is in semi-log scale where x-axis is logarithmic

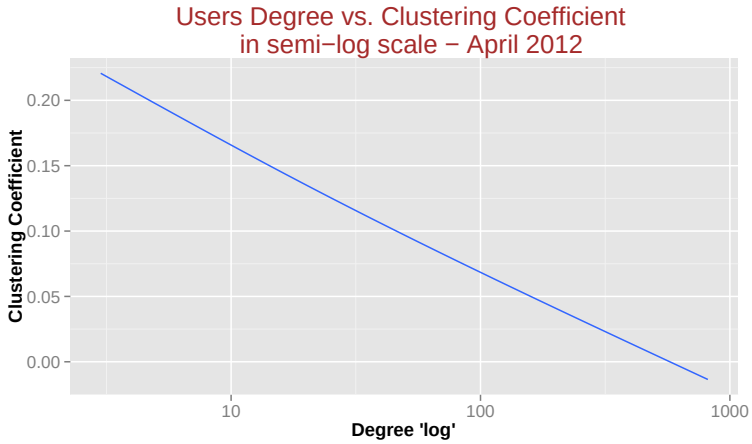

Figure 5.8: Clustering Coefficient of users with different degrees in APRIL 2012. The users with few friends are tightly clustered while the users with many friends are less clustered. The graph is in semi-log scale where x-axis is logarithmic

Figure 5.9 shows the average clustering coefficient plotted as a function of the nodes' degree for the network in July 2012. The users with low degree are tightly clustered in the network; however, the users with higher degree are tightly clustered too.

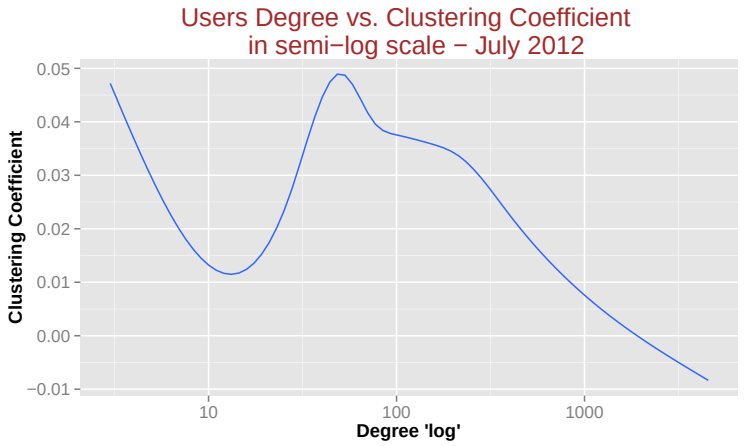

Figure 5.9: Clustering Coefficient of users with different degrees in JULY 2012. The users with very few friends are tightly clustered, and the users with many friends are tightly clustered too. The graph is in semi-log scale where x-axis is logarithmic

Beside this pattern between the average clustering coefficient against the nodes' degree, through the 11 months, the network has a small diameter and a small average shortest path, which indicates the existence of the small-world property in the Socl network [24]. Moreover, the power law fitted to the degree distribution indicates the existence of the scale-free property in the Socl network[17].

5.2 Analysis of the Interests Network Structure

5.2.1 General Structure of Interests Network Analysis

In this section, we will explore the interests of the users in the 'Socl' Network. We have about 23000 total number of interests in our dataset that users followed between January and November 2012. These interests vary on their number of followers in a long tailed distribution as shown in figure 5.10.

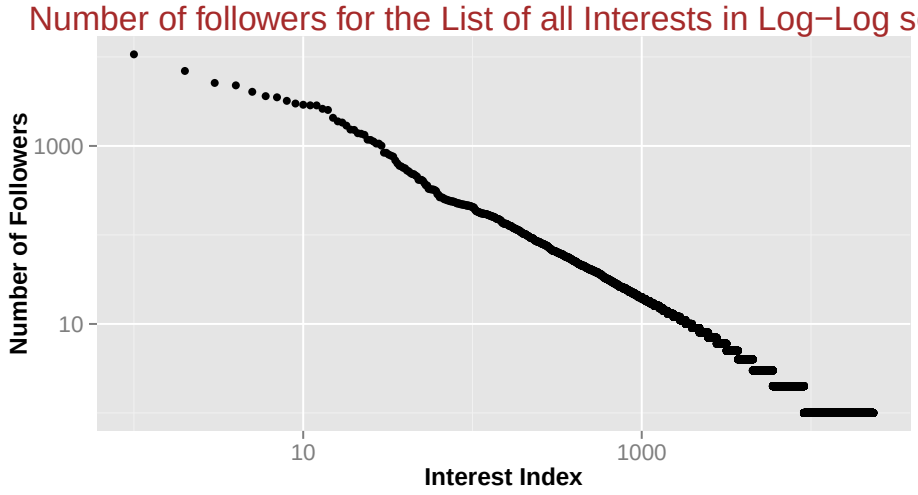

Figure 5.10: The scatterplot of the Number of followers Distribution for the list of All interests existed between January to November 2012. The graph is in log-log scale where both of x-axis and y-axis are logarithmic

Table 5.7 shows the number of the interests in three different ranges of followers number. We see that the biggest portion of the interests have less than 10

followers, which may not as interesting to consider in this analysis.

Range of No. of Followers	Approximate No. of Interests
less than 10	22376
Between 10 and 99	824
Between 100 and 999	52
more than 1000	18

Table 5.7: Summary of No. of Followers in the list of all Interests

We build an undirected weighted network to represent the Interests network. Nodes in the network represent the interests, and the weight of the edge between any two nodes represents the number of users who followed both the source and the target interests.

For a clear insight in the interests network, figure 5.11 shows a cloud representation of the interests/tags network. The size of the nodes represents their number of followers, and their color shows the communities among the interests. The communities of Interests represent the relevant and similar interests which have common users in between. Interestingly, similar or related interests are clustered together which indicates that users tend to have interests in similar topics or fields. For example, interests such as art, photography and architecture are grouped in one community and it makes sense that user who is interested in one of them is interested in the other two as well.

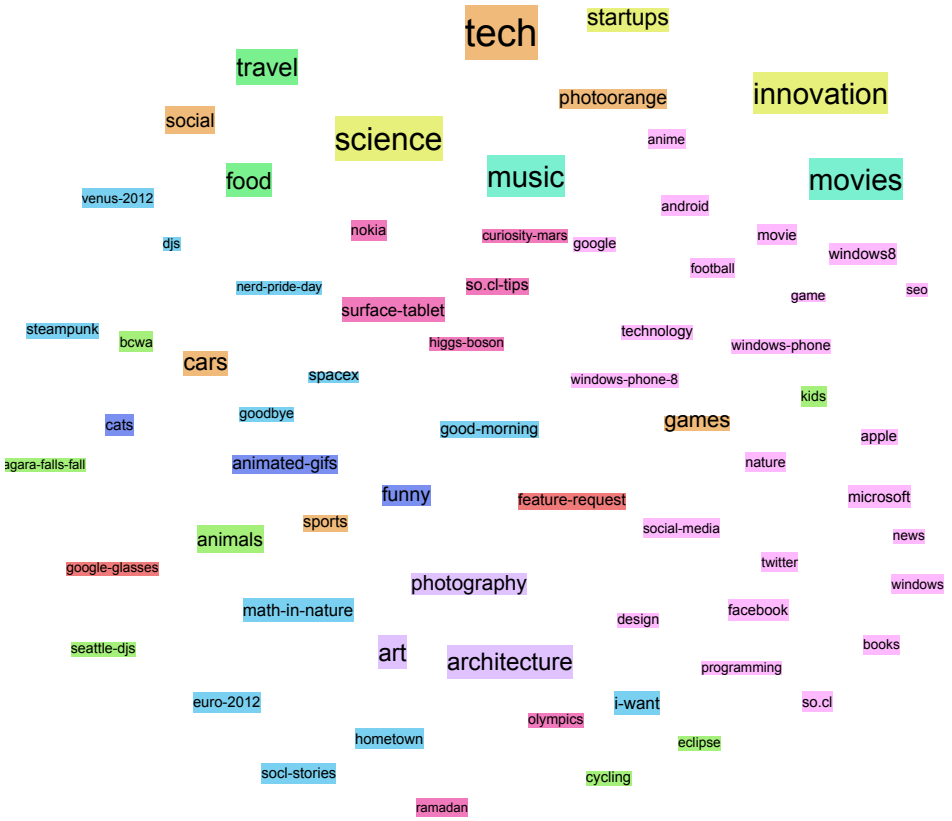

Figure 5.11: The cloud tags of the interests that have 100 and more followers between the January and November 2012. The size of the nodes represent their number of followers while their color represents the communities. The communities in the Interests network represent the relevant and similar interests which have common users in between

5.2.2 Fitting the Number of Followers for All Interests to Power-Law

We will follow the same procedure in section 5.1.3 for fitting a power law to the distribution of the number of followers for all Interests.

- **STEP1:Estimating the Power Law Parameters**: first we estimated the power law parameters x_{min} and α as 2 and 2.1 respectively.
- **STEP2: Goodness of Fit tests**: Using the goodness of fit test, we have the values of $p-value$ and goodness of fit as 0.0288 and 0.0124 respectively. Considering the significance level is 0.1, the p-value is less than the significance level and therefore the power law is not a plausible hypothesis in this case. We reject the null hypothesis and we conclude that the distribution of the number of followers for Interests and the distribution of power law data don't came from same distribution.
- **STEP3: Alternative Distribution**: Using likelihood ratio test, we compared the power law to the log-normal distribution. The value of LR equals to -0.32 which is close to zero and such insufficient value can't be used to decide on which distribution fits better, since this decision could be affected by the statistical fluctuation.

Figure 5.12 shows the CDF plot for the number of followers distributions for all interests, with a red power law fit line.

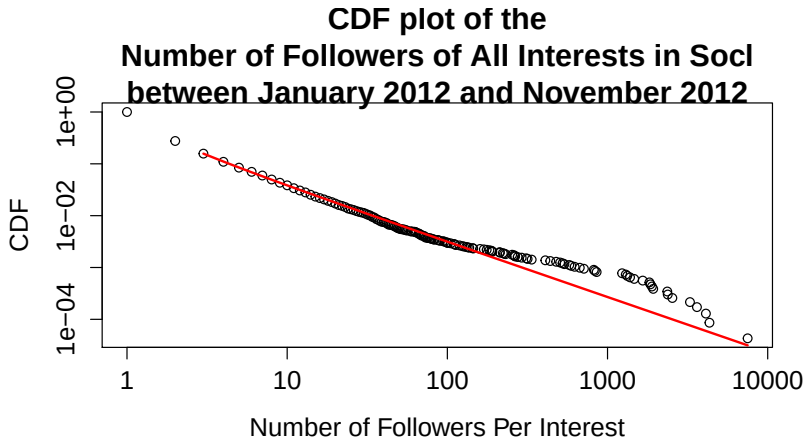

Figure 5.12: The CDF plot of the Number of Followers for all interests between January and November 2012, with power low fit line

5.2.3 Correlation between the Users' Degree and Interests Number

The number of the interests that the user follows is positively correlated with his number of followers with correlation coefficient equals to 0.59. Moreover, the number of the user's interests is positively but weakly correlated with the number of his followees with correlation coefficient equals to 0.41. This suggests that the number of followers and followees of a user is strongly related to his number of interests. This means, the increase of the number of followers or followees of a user tend to associate with increase in his number of interests and similarly the decrease of the number of followers or followees of a user tend to associate with decrease in his number of interests.

Beside considering the strong positive correlation between number of followers and followees in section 5.1.3. We may explain these correlations in simple scenarios such as when the user follow new users who could have different interests, which he will adopt later and this is called *Contagion*. Another explanation could be, when the user has an interest, he will expand his friendship by following people with similar interests as his own, and therefore, he gets more friends who

have similar interests as his own and this is called *Homophily*.

In the next section we will explore more the homophily and the contagion effect in the network to extend our understanding to the relationship between the users' interests and their friendship behavior.

5.3 Analysis of the Homophily and Contagion in shaping the users friendship and interests

To investigate the homophily principal in the active users network, we compared two probabilities for every user based on his friendship and interests. The first probability evaluates to which extend the user will be affected by his friends who have different interests than his own, and if he will adopt these interests after them which indicates the *contagion* principal. The second probability examines to which extend the user will be making friendship with others who have similar interests as his own which indicates the *homophily*.

These two probabilities have a strong negative relationship with correlation coefficient equals -0.41. This strong correlation can indicate that users are tend to have strict behavior, either to be affected by the friends of different interests *more contagious* or to be affected by these with similar interests *more homophilic*.

Figure 5.13 shows the probability density functions and the histogram of both probabilities combined in the same graph, where the homophily represented in red and the contagion in blue. The histograms of both probabilities are plotted together, which resulted on the shaded area in purple color.

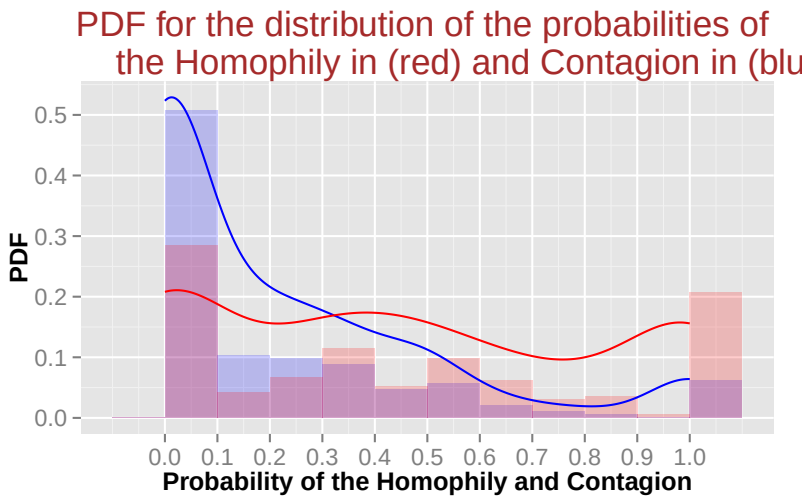

Figure 5.13: Probability Density Function and Histogram of both: The *probability of the users to adopt new interests similar to that of their friends contagion* in blue and the *probability of the user to get new friends of similar interests as his own homophily* in red

The x-axis corresponds to the values of both probabilities in the range from 0 to 1, while the y-axis corresponds to the probability density function.

The contagion probability, presented with the blue line in the graph, shows that about 44.5% of the users have zero probability of being affected by their friends' interests which means about almost the half of the users do not follow their interests after their friends. The rest of the users, 55.4% of them has been adopting new interests after their friends with a varying probability, however, only 6% of the users have probability of 1 to be affected by their friends to adopt new interests. The mean value is 0.211, indicates that on average if we randomly selected a user the probability of him being adopting new interests after his friends equals to 21.1%. Knowing that about 44.5% of the users which is almost the half were not actually adopting new interests after their friends at all, then 21.1% probability could indicate the low presence of heterophily where users get affected by friends with different interests which they adopt later.

The homophily probability, presented with the red line in the graph, shows that about 27.5% of the users have zero probability to get new friends with similar interests as their own, and about 20.7% with probability equals 1. The rest of the users, 51.8% of the users, are distributed in a similar to normal curve around the average value 0.445, indicates that on average if we randomly selected a user the probability of him being gaining new friends with similar interests as his own equals to 44.5%. Knowing that the almost the half of users here were distributed almost to equal portions in the extremes, where the rest of users half are about normally distributed around the mean.

Knowing that the half of users here were distributed almost to equal portions in the extremes, where the other half are about normally distributed around the mean, then 44.5% probability could indicate the high presence of homophily where users get new friends with similar interests as their own compared to the contagion. Table 5.8 shows a brief comparison between the homophily and contagion in terms of their mean probability, probability of them to not occur at all (probability=0) and the probability of them to absolutely occur (probability=1).

Prob	Homophily	Contagion
0%	27.5%	44.5%
100%	20.7%	6%
mean	44.5%	21.1%

Table 5.8: comparison between the homophily and contagion in terms of their mean probability, probability of them to not occur at all (probability=0) and the probability of them to absolutely occur (probability=1)

CHAPTER 6

Conclusion

6.1 Findings and Contribution

In this thesis, we conducted an in-depth analysis in the *So.Cl* network as an example of the interest-based social networks. We analyzed the structure of the evolving network with the changes over a time interval of about one year. We examined some of the social network features and theories in the users behavior regarding friendship and interests. Furthermore, this study observed the homophily behavior of the users in *So.Cl* network given their interests and their friendship behavior. It investigated the relation between the friendship behavior of the users and their interests and to which extend each of these affect the other.

The main contribution of this thesis is to extend the literature of the experimental studies in the online social networks to verify the proposed theories and models of the social networks in terms of the dynamics and features of the social networks.

This thesis provided a measurement study and analysis on the dynamics of the

topological structure and the features of *So.Cl* network. We presented an insight into the dynamic change, where we observed interesting evolutionary patterns on the structure of the 'Socl' network as it went from a private project within Microsoft corporation to a project shared with three universities and invitation only stage before it released to the public. This network grew while going through different phases that are significantly different in terms of size and users' characteristics.

Beside fitting the degree of the users to power law, we noticed the strong positive correlation between the in-degree and out-degree corresponding to the number of the followers and the number of followees. Moreover, we observed the positive correlation between the friendship of the users and the number of the interests they have. Additionally, we investigated the principals of homophily and contagion in the network by testing the probability of the users to gain friends with similar interests of their own or to adopt the interests of his friends.

6.2 Future Work

As in this work we studied the structures of the users network and the interests network separately. In future, we want to extend the analysis of the Interests network and apply the same dynamic aspects we applied to the users network while considering the friendship between their followers. We want to involve the users behavior and observe the effect of different friendship patterns in the changes in the interests network. By visualizing the Interests network given the friendship between their followers, we may find interesting patterns that affects the structure of the Interests network over time.

CHAPTER 7

Abbreviations

FTP File Transfer Protocol

OSN Online Social Networks

LR Likelihood Ratio

Bibliography

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[4] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[5] comScore. comscore releases february 2014 u.s. search engine rankings, 2014. URL https://www.comscore.com/Insights/Press_Releases/2014/3/comScore_Releases_February_2014_U.S._Search_Engine_Rankings. [Online; accessed 16-March-2014].

[6] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

[7] Cedric Dupont. Searchwiki: make search your own. http://googleblog.blogspot.ae/2008/11/searchwiki-make-search-your-own.html, 2008.

[8] Cedric Dupont. Stars make search more personal. http://googleblog.blogspot.ae/2010/03/stars-make-search-more-personal.html, 2010.

[9] Brynn M Evans and Ed H Chi. Towards a model of understanding social search. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 485–494. ACM, 2008.

[10] Colin S Gillespie. *Fitting heavy tailed distributions: the poweRlaw package*, 2013. R package version 0.20.1.

[11] Mathieu Jacomy, Sebastien Heymann, Tommaso Venturini, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization. *Medialab center of research*, 2011.

[12] Bernard J. Jansen, Brian Keith Smith, and Danielle L. Booth. Viewing online searching within a learning paradigm. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR*, pages 859–860. ACM, 2007. ISBN 978-1-59593-597-7.

[13] Sep Kamvar. Search gets personal. http://googleblog.blogspot.com/2005/06/search-gets-personal.html, 2005.

[14] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[15] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, 407(1):458–473, 2008.

[16] Marissa Mayer. Universal search: The best answer is still the best answer. http://googleblog.blogspot.ae/2007/05/universal-search-best-answer-is-still.html, 2007.

[17] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[18] Meredith Ringel Morris. A survey of collaborative web search practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1657–1660. ACM, 2008.

[19] Meredith Ringel Morris. Collaborative search revisited. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1181–1192. ACM, 2013.

[20] Wikipedia. Web search engines, wikipedia the free encyclopedia, 2013. URL http://en.wikipedia.org/wiki/Centrality. [Online; accessed 16-March-2014].

[21] Wikipedia. Web search engines, wikipedia the free encyclopedia, 2013. URL http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient. [Online; accessed 24-March-2014].

[22] Wikipedia. Web search engines, wikipedia the free encyclopedia, 2013. URL http://en.wikipedia.org/wiki/Web_search_engine. [Online; accessed 16-December-2013].

[23] Wikipedia. Web search engines, wikipedia the free encyclopedia, 2014. URL http://en.wikipedia.org/wiki/Power_law. [Online; accessed 24-March-2014].

[24] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.