

HM  
743  
F33  
P38  
2010

# CHARACTERIZATION OF USER NETWORKS IN FACEBOOK

by

Fatemeh Pakzad

B.Sc. Applied Mathematics, Amirkabir University of Technology, 2002

A thesis

presented to Ryerson University

in partial fulfillment of the

requirement for the degree of

Master of Science

in the Program of

Computer Science

Toronto, Ontario, Canada, 2010


© Pakzad, Fatemeh 2010

PROPERTY OF  
RYERSON UNIVERSITY LIBRARY

## Declaration


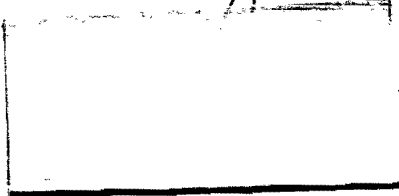
I hereby declare that I am the sole author of this thesis.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

---

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in parts, at the request of other institutions or individuals for the purpose of scholarly research.

# CHARACTERIZATION OF USER NETWORKS IN FACEBOOK

Fatemeh Pakzad

M.Sc. Computer Science, Ryerson University, 2010

## Abstract

The purpose of this thesis is to illuminate several characteristics of Facebook user networks in order to model user network formation in Facebook. Based on the amount of data gathered during a seven-month period, we conclude that the user network node degree distribution in Facebook follows a Lognormal distribution, and the distribution of the increase in the number of fans of the Facebook's web pages follows the Weibull distribution. Also we present that the Facebook user network graph has small world characteristic. Finally, by using the distributions that modeled the Facebook user networks with the small world characteristic, this thesis proposes a new algorithm in order to simulate a Facebook user network graph and the increase in the number of Facebook's web page fans. We built the software that generates the graphs with similar static and dynamic characteristics of Facebook user networks.

## **Acknowledgments**

I owe my deepest gratitude to my supervisor, Dr. Abdolreza Abhari, whose encouragement, guidance and support from the initial to the final level has enabled me to develop an understanding of the subject.

I would also thank my family who helped me in their own way to achieve my goal.

# TABLE OF CONTENTS

## Chapter 1

<b>1. Introduction .....</b>	<b>1</b>
1.1 Thesis Motivation .....	3
1.2 Thesis Objectives and Scope.....	4
1.3 Research Contributions .....	4
1.4 Thesis Outline .....	5

## Chapter 2

<b>2. Background Information and Related Works .....</b>	<b>6</b>
2.1 Background Information .....	6
2.1.1 Web 2.0 .....	6
2.1.2 Online Social Networks .....	7
2.1.2.1 Facebook .....	7
2.1.2.2 Other Online social Networks.....	9
2.1.3 Statistical Definitions .....	12
2.1.4 User Network Graph Definitions .....	15
2.1.4.1 Small world Phenomenon .....	15
2.1.4.1.1 Average Clustering Coefficient .....	15
2.1.4.1.2 Average Short Path Length .....	17

2.1.5 Crawling Methods in Online Social Network.....	17
2.2 Related Works .....	18
2.2.1 Online Social Networks Workload Characteristics .....	18
2.2.1.1 Facebook Workload Characteristics .....	18
2.2.1.2 YouTube (OSN) Workload Characteristics .....	20
2.2.1.3 Other Online Social Networks Workload Characteristics ...	23
2.2.2 Study the Small World Characteristic in Online Social Networks ...	24
2.2.2.1 The Small World in YouTube .....	25
2.2.2.2 The Small World in Other Online Social Networks .....	26

## Chapter 3

<b>3. Methodology and Measurements .....</b>	<b>27</b>
3.1 User Network Analysis .....	28
3.2 Facebook Web Page fans .....	31
3.3 Small World Characteristic in Facebook .....	39
3.3.1 Average Clustering Coefficient .....	40
3.3.2 Average Short Path Length .....	41
3.3.3 Discussion of the Existence of Small World Characteristic .....	42
3.4 Conclusions .....	44

## Chapter 4

<b>4. Synthetic Workload Simulator .....</b>	<b>46</b>
4.1 Generating Facebook User Network Graph.....	46

4.1.1 Generate the Node Degree distribution .....	46
4.1.2 Generate the Graph .....	48
4.1.3 Generate the Small World Network Graph .....	49
4.1.3.1 Optimizing the Average Clustering Coefficient Value.....	49
4.1.3.2 Average Short Path length .....	52
4.2 Simulation Graph.....	53
4.3 Generating Web Page Fans Graph.....	55
4.4 Conclusions .....	59

## **Chapter 5**

<b>5. Conclusions and Future Works .....</b>	<b>60</b>
5.1 Conclusions .....	60
5.2 Future Works .....	62
References .....	64

## LIST OF TABLES

Table 3.1	Collected data sets .....	27
Table 3.2	Statistics of the node degree of Facebook .....	30
Table 3.3	Kolmogorov-Smirnov Test	
	Node degree distribution .....	30
Table 3.4	Summary statistics	
	The increase in the number of fans in the four web page fans .....	35
Table 3.5	Kolmogorov-Smirnov Test	
	The increase in the number of fans in four web pages fans .....	36
Table 3.6	Summary statistics	
	The increase in the number of fans in four web pages fans .....	38
Table 3.7	Kolmogorov-Smirnov Test	
	The increase in the number of fans in four web pages fans .....	39
Table 3.8	The comparison table	
	Facebook user network and Erdos-Renyi random graph .....	44
Table 4.1	Summary statistics of the generated data .....	55
Table 4.2	Kolmogorov-Smirnov Test	
	The generated data.....	56



## LIST OF FIGURES

Figure 2.1	The sample graph .....	16
Figure 3.1	The node degree distribution of Facebook .....	31
Figure 3.2	The distribution of increase in the number of Starbucks web page fans .....	32
Figure 3.3	The distribution of increase in the number of CNN web page fans .....	33
Figure 3.4	The distribution of increase in the number of MTV web page fans .....	33
Figure 3.5	The distribution of increase in the number of Facebook web page fans.....	34
Figure 3.6	The distribution of increase in the number of Timhortons web page fans .....	36
Figure 3.7	The distribution of increase in the number of Friends TV show web page fans..	37
Figure 3.8	The distribution of increase in the number of Oprah TV show web page fans....	37
Figure 3.9	The distribution of increase in the number of a political person web page fans...	38
Figure 4.1	Generating Graph form.....	47
Figure 4.2	Generating Network Graph form.....	48
Figure 4.3	User Facebook Network Simulation Result form.....	52
Figure 4.4	Simulated Facebook user network graph for 50 nodes .....	53
Figure 4.5	Simulated Facebook user network graph for 100 nodes.....	54
Figure 4.6	The distribution of the generated data .....	55
Figure 4.7	Generating Graph form .....	57
Figure 4.8	Generating Web Page Fans form.....	57
Figure 4.9	The increase in the number of web page fans during the simulation of 200 days.	58

## LIST OF ALGORITHMS

Algorithm 3.1	Calculation of the average clustering coefficient in a graph .....40
Algorithm 3.2	Generating Erdos-Renyi Random Graph .....43
Algorithm 4.1	Optimize the average clustering coefficient .....50

## LIST OF ACRONYMS

### ACRONYMS

### DEFINITIONS

<b>WWW</b>	World Wide Web
<b>BFS</b>	Breadth First Search
<b>OSN</b>	Online Social Networks
<b>REST</b>	Representational state transfer
<b>CV</b>	Coefficient of Variation
<b>PDF</b>	Probability Density Function
<b>CDF</b>	Cumulative Distribution Function
<b>WCC</b>	Weakly Connected Component: The subgraph, in which any two nodes are reachable to each other
<b>CCDF</b>	Complementary Cumulative Distribution Function
<b>SLA</b>	Service Level Agreement
<b>LWL</b>	Least Workload Left: It is a caching algorithm that sends the request to the server who has the least remaining workloads.
<b>QoS</b>	Quality of Service
<b>FCFS</b>	First Come First Served
<b>K-S test</b>	Kolmogorov-Smirnov test



# Chapter 1

## 1. Introduction

The establishment of web 2.0 (the new generation of World Wide Web (WWW)) was a big development in the field of computer technology. A key feature of web 2.0 is the introduction of web sites that enable the users to build social networks. Online social networks allow users to share and organize content, videos, photos, and even their comments. The famous web 2.0 online social networking sites are Facebook, Twitter, Orkut, Flickr, LiveJournal, and YouTube. One of the most popular social networking sites is Facebook.

One of the advantages of these web sites is helping market producers to boost their productivity. For instance, on Facebook, many companies such as Starbucks have a web page. Therefore, the market producers can easily contact their customers, listen to their comments and feedback and learn from them. They are also able to distribute their company's information very fast. By using web 2.0 technology, marketing companies can contact their customers or broadcast their promotional events without spending a lot of money on commercial advertising on TV or radio. In today's difficult economic times, marketing companies are looking for new ways to introduce their market strategies. Therefore, by using social networks such as YouTube, Facebook and Twitter, market producers can introduce their brands without spending a large amount of money.

As mentioned in [1, 2], web 2.0 technology allows users to share and upload content within Internet communities. Web 2.0 has changed the role of the users on the Internet. By Web

2.0, the users are not only consumers, but also part of the information source. It allows users to share their opinions via blogging sites. Furthermore, it allows them to share video clips and photos through specific web sites such as YouTube and Facebook.

Due to the popularity of online social networking sites, finding the characterization of user networks in online social networking sites has become a new research topic. Based on the online social networking site, various kinds of user networks can be defined. For example, in the YouTube user network, two users can link to each other by watching each other's video files. In this research for understanding the Facebook social network, we consider two user networks in Facebook, and we model both of these user networks. The first model is related to the user networks (i.e., user profiles and their friends) which we refer to as static modelling. By using user network, two users in Facebook can link to each other through their friendship link. In this research we have mathematically modeled 20,019 Facebook users with the number of their friends. By modelling the user networks in Facebook synthetic workload simulators can be built that can be used for improving content delivery for Facebook users which can reduce traffic both in the server and network sites.

The second user network study that we consider in this work is referred to as dynamic modelling, which is related to the increase in the number of web page fans for eight popular web pages in Facebook. In this research the increase in the number of fans for eight web pages during 131 days is modeled. Analyzing social networking will help us to identify how much fast companies can broadcast their events by using Facebook.

We also study the existence of the small world phenomenon in Facebook. The existence of the small world phenomenon in the user networks shows that there are many paths between

two nodes of the user network graph. Thus, news can spread faster between the users. At the end of this thesis, by applying the Facebook user network features that are analyzed in this thesis, the Facebook user network graph is simulated. The distribution of increase in the number of Facebook web pages fan which is analyzed at the beginning of this thesis is used to simulate the increase in the number of users in the Facebook web pages.

The remainder of this chapter is organized as follows: Thesis motivation is in section 1.1. The objectives and scope of the thesis are presented in section 1.2. The thesis contributions are discussed in section 1.3. The organization of the rest of the thesis is presented in section 1.4.

## **1.1 Thesis Motivation**

Network service providers have been faced with technical challenges that are related to the bandwidth consumption of online social networking web sites such as Facebook. Therefore, it is better to know more about characteristics of these web 2.0 sites. Several benefits can be achieved from this study. One of the implications is building the synthetic workload simulators that can be used in the experiments for evaluating the methods that address bandwidth consumption problem.

Nazir *et al.* [3] has mentioned that a small number of users generates large traffic through the applications installed on their Facebook pages. In Facebook, active users upload many videos and pictures to share them with their Facebook friends. This then generates large data on storage systems and many requests on servers of Facebook data centres. Modelling user networks leads us in building the synthetic workload simulators which measure the performance of the methods trying to reduce the traffic of the web 2.0 sites similar to Facebook.

## 1.2 Thesis Objectives and Scope

The goal of this research is to understand social network characteristics. Therefore, modelling user networks in Facebook is the main objective of this research. These findings have the capability of helping researchers determine the structure and speed of data transfer in user networks. To pursue this goal, first the user networks graph in Facebook should be analyzed. Then other graph characteristics such as the average short path length and the average clustering coefficient which can be used to examine the existence of the small world characteristic in the Facebook user graph should be studied. Data collection, empirical measurement and curve fitting are the methodologies that we used for this study.

In this work, to model the static characteristic of a user network, a user and the number of user's friends are collected. We found that the increase in the number of the user's friends changed very slowly during the interval of our data collection.

To model the dynamic characteristics of the user network, the increase in the number of web page fans was studied because it is changing frequently. We did not do any traffic performance measurement test with the simulator because it is considered as our future work.

Another limitation of this work is the study of only these two conditions (small world network characteristic and user network graph) to simulate online social networks. However, there are more characteristics for online social networking web sites that were not in the scope of this work.

## 1.3 Research Contributions

This thesis has the following main contributions:

- Modelling Facebook user network node degree which fits the Lognormal distribution



- Modelling the increase in the number of web pages fans that follow the Weibull distribution
- Determining the existence of small world phenomenon in Facebook user networks
- Using the above models to build the synthetic workload generator for simulating static and dynamic characteristics of user networks

## **1.4 Thesis Outline**

The remainder of this thesis is organized as follows: There is an overview of the online social networking features and small world phenomenon concepts along with related work on online social networking characteristics in chapter 2. Chapter 3 presents the methodology and measurement. The implementation is presented in chapter 4. Finally, Chapter 5 concludes our work.

## **Chapter 2**

### **2. Background Information and Related Work**

In this chapter, the required background information of this thesis along with the similar works in the study of social networking characteristics are presented. Background information is presented in section 2.1 while section 2.2 presents all related works.

#### **2.1 Background Information**

This section presents the required background information that is related to the thesis. The web 2.0 background information is presented in section 2.1.1. The online social network's background information is discussed in section 2.1.2. The statistical definitions used in the thesis are presented in section 2.1.3; the user network graph definitions are presented in section 2.1.4. Finally, section 2.1.5 explains a few crawling methods that can be used for graph data measurement.

##### **2.1.1 Web 2.0**

As mentioned in [1], web 2.0 technology allows users to share and upload content within Internet communities. Web 2.0 has changed the role of Internet users. Through web 2.0, users are not only consumers, but also part of the information source. It allows users to share their opinions via blogging sites. Furthermore, it allows them to share video clips and photos through specific web sites such as YouTube and Facebook. The technology behind web 2.0 is Ajax which merges JavaScript and XML with traditional HTML. Thus, servers should be able to store and transfer a large amount of data to a large number of users.

To summarize, web 2.0 allows client-side applications to migrate from Personal Computers (PC) to server-side applications [1]. This journey to the server-side has several advantages. For instance, services can be delivered to users as well as businesses companies. It also has a loose coupling feature. Loose coupling allows for different environments to deliver service to users. Web 2.0's loose coupling feature is one that is complicated. Thus, it is called server-side mashups. Server-side mashups allows a server to receive information from different sources and present it as a single point to consumers.

In web 2.0, different technologies such as AJAX, Representational state transfer (REST), and Flash are all examples of popular technology used for the purpose of delivering video contents and to drive web 2.0 to clients. Some of the emerging applications in web 2.0 are listed as follows: social networking, server-side productivity application and software-as-a- service.

One of the challenges in web 2.0 is to solve its scalability problem. Recently, new distribution software for memory caching called Memcached, Gear6 [4], has been introduced. The Gear6 solution improves scalability of web application and quality of service. Gear6 also reduces infrastructure cost.

### **2.1.2 Online social Networks**

As previously mentioned, Facebook is evidently one of the most popular online social networks. The background information required for understanding Facebook is presented in section 2.1.2.1. Section 2.1.2.2 presents background information related to the other online social networks.

### 2.1.2.1 Facebook

One of the most popular Online Social Networks (OSN) is Facebook which has more than 500 million users [5]. In Facebook, each user has some abilities such as: adding friends, updating their profile, joining different networks, sending and receiving private messages, and writing comments on photos, videos, and their friends' status. There are several networks including universities, cities and countries that can be formed by the users through Facebook.

According to [6], the key feature of Facebook is an application platform which lets a third party social networking to be developed. In Facebook, when a user wants to use an application, it should install the application in its profile [7]. In Facebook, the popularity of application is defined as a number of application's users, thus the older applications have more chance to become popular.

One of the features of Facebook which is referred to as "activity network" is defined as an interaction between users [5]. In Facebook, while some user pairs have strong activity networks, others have weak activity network.

#### *Facebook Applications:*

They are several kinds of applications in Facebook as defined in [3]:

- *Social gaming:* Some applications let users to play games with other users an example of which would be Fighters' Club (FC).
- *Non gaming application:* There are many non gaming applications on Facebook such as: Got Love (GL) and Hugged which are used between friends.

The top 7 categories of non gaming application are listed as follow:

- *Friend comparison* which allows users to indicate their top friends.

- *Casual communication* that lets users to exchange messages.
- *Rating taste matching and recommendations* that enable users to rate music, restaurants and etc.
- *Gestures* which lets the users to show their gestures over poking and other motions.
- *Self expression* that lets users to express their mood.
- *Gifting* which enable users to exchange virtual gifts with their friends.
- *Meeting people* that is kind of online meeting

#### *Web Pages in Facebook:*

There are many web pages in Facebook developed for different TV shows, political people, celebrities, public places, and market companies. The users can become fan of these web pages, get related news and share news, or comments the other fans as well. Each web page in Facebook is same as user's page, but they have fans instead of friends. Most of these web pages are public for everyone even for people who are not Facebook's member.

#### **2.1.2.2 Other Online Social Networks**

- *Twitter* was established in October 2006. It is a short message service which allows users to store and forward messages to their friends.
- *Orkut* [8] is an online social network which allows for its users to find friends and share video clips and photos.
- *LiveJournal* [9] allows its users to share content and provide comments for other users.
- *Flickr* [10] is a photo sharing online social network.

### *YouTube:*

YouTube was established in 2005 [11]. It has made a huge difference in Internet. The video sharing allows users to upload videos and to tag videos in their blogs or web pages. It also lets YouTube users to write comments or to rate videos. One of the biggest differences between YouTube and traditional multimedia servers is the length of videos. In traditional multimedia servers, the length of each video is about 1-2 hours [11]. However, research shows that in YouTube, 97.9% of videos are less than 600 seconds. Almost 99% of videos are less than 700 seconds due to the fact that YouTube has set a ten minute limitation on each video. Before March 2006, there was no time limit on videos, so currently there are some video files on YouTube that are longer than 10 minutes in length. Currently, only a small group of authorized uploaders can upload longer videos. Most videos belong to three categories based on their length. The first category consists of videos that are less than one minute. This category contains more than 20.6% of videos and they are entertainment related. The second category includes videos with the duration length of three to four minutes. On YouTube, this consists of 17.1% of the videos their topics generally being music related. The third category includes videos that are ten minutes long. This category consists of longer videos that have been divided into 10 minute segments. According to [12], the size of the video file is similar to the video's lengths. Statistics show that 98.3% of uploaded videos have less than 25 MB, while the average video size is 8.4 MB.

For crawling YouTube, only forwarded links that will be explained next can be used, so some of the nodes in Weakly Connected Component (WCC) may not include the crawling data set [13]. In YouTube, each video has some information such as: Video ID, Uploader ID, Date Added (shows the time when the video added to YouTube), Video Category, Video Length,

Number of views, Rating, Number of Rating, Number of comments, and related videos list. Uploaders can then set related video lists along with videos that share a common topic, have the same topic, description, or tags. Therefore, in YouTube, videos share a connection with each other, while in traditional multimedia servers video clips are separate and independent from one another.

There are several video categories on YouTube such as: the most popular category is music which is about 22.9% of all YouTube video files, followed by the entertainment at about 17.8%, comedy, unavailable videos which set as private or inappropriate videos, and removed videos which are removed from YouTube by the uploader or YouTube moderator.

In [14], nine characteristics are defined for user behavior on YouTube as follows: The first five characteristics are individual user's behaviors including: number of uploaded files, total number of views per user, total number of channel views per user, and network join date. The other four characteristics are social users interaction behavior which include: clustering coefficient of a node, reciprocity (the mutual subscription action between two nodes), out-degree (i.e., number of subscription made by the user), and in-degree (i.e., the number of subscription received by the user).

*Metacafe* [15]:

It is a video sharing web site from France. It is a little different from the other videos sharing web sites mentioned in the above text because its content's creators are obligated to pay for their video files once the number of views goes over a certain number [14].

### *Yahoo Video:*

It was established in June 2006. Yahoo Video is similar to other online video sharing web sites such as YouTube and Veho. Similar to YouTube, Yahoo Video lets users upload and watch their videos of choice.

### **2.1.3 Statistical Terms Definitions**

Several statistical features are measured to model the data such as mean, median, standard deviation, and distribution function. In this section, all of the definitions that are needed in this thesis are presented.

#### *Mean:*

The summation of all discrete data set divided by the number of the data set is mean [16]. For the discrete data set with n data  $x=(x_1, x_2, \dots, x_n)$ , the mean is calculated by the following formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (1.1)$$

#### *Median:*

Another statistical feature of the data set is the median. The median is the middle number of the data set [16]. For finding the median of the data set, the data can be sorted from low to high, and then the middle data is found. If the number of data in data set is even, then the median is equal to the average of two middle data.

#### *Standard Deviation:*

Another widely used statistical feature is standard deviation. It shows the deviation of data set from the mean [16]. The standard deviation is calculated by the square roots of the



variance. For the discrete data set with n data  $x=(x_1, x_2, \dots, x_n)$ , the standard deviation is calculated as follows:

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (1.2)$$

*Coefficient of Variation (CV):*

The other statistical feature is Coefficient of Variation (CV) that is calculated by the following formula [16]:

$$CV = \frac{s}{\bar{x}} \quad (1.3)$$

where  $s$  is standard deviation and  $\bar{x}$  is mean.

*Probability Density Function:*

The Probability Density Function (PDF) of a random variable calculates the probability of occurrence of the random variable in the observation space [16].

*Cumulative Distribution Function:*

The Cumulative Distribution Function (CDF) of a random variable  $x$  is defined as the probability of occurrence of the variable of  $X$  less than or equal to a value of  $x$  [16]. It is shown by  $F(x)=P\{X \leq x\}$ . It is obvious that the CDF is not decreasing between 0 and 1.

### *Lognormal distribution:*

The random variable is Lognormal distribution if the logarithm of the variable is distributed normally [16]. The Lognormal distribution is shown by  $\ln N(\mu, \sigma^2)$ , where  $\mu$  is a location and  $\sigma$  is scale.

The PDF of the Lognormal distribution is calculated by the following formula:

$$F_x(x; \mu; \sigma) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \quad x > 0 \quad (1.4)$$

where  $\mu$  is mean and  $\sigma$  is standard deviation.

### *Weibull Distribution:*

The Weibull distribution is a continuous distribution [16]. The PDF of the Weibull distribution is calculated by the following formula:

$$f(x; a; b) = \begin{cases} ab^{-a} x^{a-1} e^{-(x/b)^a} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (1.5)$$

where  $a$  is shape and  $b$  is scale parameter of the distribution.

### *Kolmogorov-Smirnov test (K-S test):*

The K-S test measures the largest vertical distance between the CDF of fitted distribution and the empirical distribution [17]. By applying the K-S test, the best match for the empirical distribution is chosen which has the smaller distance to the empirical distribution in comparison with the other distribution. We used this test to support the visual observation for the generated graphs.

#### 2.1.4 User Network Graph Definitions

In this research, the user network is modeled as a graph for which we want to find the characteristics. So some of the features of the graph are used to analyze user networks that are defined in this section.

##### 2.1.4.1 Small World Phenomenon

Small world phenomenon introduced by Milgram in [18] indicates that any two users of the network can be reachable to each other. In other words, all users have the convenience of reaching each other by merely following a small number of links. Many real world networks also have small world characteristic [19, 20, 21]. Pool *et al.* [22] discovered the small world characteristic in their work in human contact net as well. Small world network is identified by two properties: a high clustering coefficient and a short path length. The average clustering coefficient is explained in section 2.1.5.1.1 followed by the definition of the average short path length of the graph in section 2.1.5.1.2.

##### 2.1.4.1.1 Average Clustering Coefficient

One of the features of the social network graph is a clustering coefficient. The Clustering coefficient indicates that there is a higher chance for two people who have a mutual friend to be friends than two strange people [23, 24]. For example, if there is a person A and this person is friend with the people B and C, also if there are existing two strange people D and E, the probability of existence of friendship between B and C is much higher than the probability of existence of friendship between D and E. Usually the clustering coefficient is presented by the formula that explained below:

$$C_A = N_{FA} / N_{PA} \quad (1.6)$$

where  $C_A$  presents the clustering coefficient of node A in figure 2.1 and NFA is the number of all edges between node A's Friends and NPA is the number of all the possible edges between node A's friends.

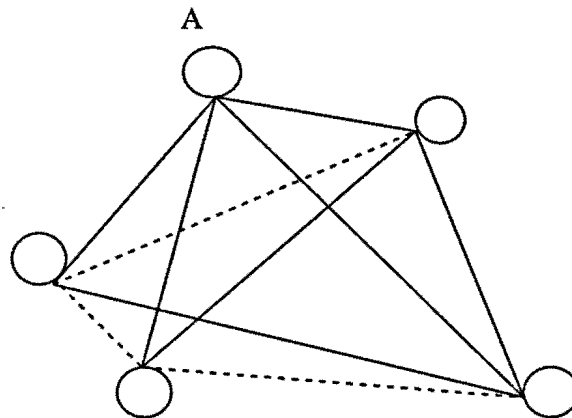


Figure 2.1 The sample graph

As it is shown in Figure 2.1, node A has four neighbours. The numbers of all possible edges between the node A's neighbours without considering A is 6; and the actual number of edges between the node A's neighbours (excluding itself) is 3, so according to formula 1.6, the clustering coefficient of node A is:

$$C_A = 3/6 = 0.5$$

As we know, if node A has  $k$  neighbours, the number of all possible edges between node A's neighbours is  ${}_kC_2$  or  $k!/2!(k-2)!$  which is equal to  $k(k-1)/2$ . Therefore, Formula 1.6 is changed to Formula 1.7 that we used in the proposed algorithm.

$$C_A = 2NFA / k(k-1) \quad (1.7)$$

Average clustering coefficient of the graph is defined as the average of clustering coefficient of all graph nodes.

#### **2.1.4.1.2 Average Short Path Length**

The other feature of the graph that is needed to identify small world characteristic is the average short path length. The short path length between two nodes presents the minimum number of hops that link those two nodes [25]. The average short path length of a node indicates the average number of short path length between that node and all other nodes of the graph. To examine the existence of small world characteristic, the average short path length of the graph is needed, which is the average short path length of all the graph nodes.

#### **2.1.5 Crawling Methods in Online Social Networking**

They are several common ways to crawl a data set from different web sites such as Orkut, LiveJournal, Flickr, MetaCafe, YouTube, and Facebook.

##### *Forwarded Link:*

It is started from a node in a directed graph and the graph is crawled by following the directed forwarded link [14]. However, using forwarded link does not tend to crawl the whole WCC.

##### *Reverse Link:*

It is started from a node in a directed graph and the graph is crawled by using both forwarded and back warded link and it helps to crawl the whole WCC [14].

### *Breadth-First Search (BFS):*

It is a graph search algorithm which starts from a random node. Then, all its neighbours are added to the data list followed by a crawl of all the neighbours of its neighbours until sufficient data is gathered [26].

## **2.2 Related Works**

As mentioned before, Facebook has become one of the most well-known online social networks. Therefore, there are several works in this area that are presented in this section. Section 2.2.1 presents some of the related works on different online social networks characterization, and section 2.2.2 presents the related works on studying the small world characteristic in online social networks.

### **2.2.1 Online Social Networks Workload Characterization**

In this section, related works on different online social networks workload characterization is presented. At first, section 2.2.1.1 presents the Facebook workload characteristics, followed by YouTube workload characteristics (i.e., the largest online social network) in section 2.2.1.2. Finally the characterization for the other online social networks will be presented in section 2.2.1.3.

#### **2.2.1.1 Facebook Workload Characteristics**

Nazir *et al.* [3] studied three different well-known Facebook applications to gather data set for user network workload characterization. These applications are FC (i.e., Fighters' Club), GL (i.e., Got Love), and Hugged which can be installed on Facebook user's profile and they are used by more than 8 million users. Facebook usage decreases on weekends and holidays. Different usage characteristics are analyzed such as geographical distribution of users, user

interaction and response time. The result shows that most of these applications' are installed by the users from the USA, UK and Canada. Also, the study shows the small number of users generates large traffic through the applications installed on their Facebook page. These users have daily usage of these applications. The response time is defined as the number of seconds passed between sending an application request from one user and response to the request from the target user. In Facebook, the average response time is 14.8 hours for requests between friends who are located in different countries with different time zone and 15.1 hours for the local requests. Unlike GL and Hugged in which a user has interaction only with the friends inside the network, FC community structure is more expanded because in FC application users have a chance to interact with strangers as well as their friends.

Viswanath *et al.* [5] evaluate the user interaction on Facebook at both microscopic and macroscopic level. In other words, user interaction is assessed precisely, and its impact on the structure of the activity network is evaluated. Two data sets are collected from New Orleans Network: the first one is collecting the users' names by using BFS (i.e., Breadth First Search) and the second data set is collecting wall posted for all the users whom are captured in the first data set. The majority of the user pairs have the weak activity network. This crawling has some limitation as well: first, in Facebook, they are several ways to interact with other users such as messages, comments on photos, videos, and wall posting. However, in this research only wall posting is considered in the data set. Secondly, the data set only contains the users' profile from the New Orleans network, so the data set is much smaller than the whole Facebook network.

The data set is divided into two parts. The first group which contains the majority of the users and it includes the users pairs with lower than 5 wall postings. The second group contains the users' pairs with more than 5 wall postings. Analyzing the first group data set indicates that

the first interaction between these users' pairs took more than a month from the time that the links (friendship) are established. Often the birthday greeting is the first wall posting for the first group. On the other hand, in the second group, the first interactions happen after a few days from the establishment of the links between the users. However, the number of wall posting decreases over the time. Therefore, the microscopic level of interaction changes rapidly over the time. While, the macroscopic features such as an average node degree, clustering coefficient, and average path length are stable over the time.

Gjoka *et al.* [6] analyzed different Facebook application data over six months. Two data sets are collected: the first one is the number of applications installed and number of daily active users for each application, and the second data set captures data from users' profiles over a week. The first data set is more general than the second one. The number of installations grows over the time; however, the number of daily active users decreases over the time. The distribution of the number of daily active users of considered applications follows power-law distribution. As it is mentioned above the first data set is gathered over six-month period while the second data set is collected over a week. Therefore, the first data set is more general than the second data set. However, the results of both data sets show that in spite the fact of difference between data sets sizes, the result of the second data set follows the first data set result. Therefore, it is as reliable as the other one.

#### **2.2.1.2 YouTube (OSN) Workload Characteristics**

As mentioned before, the Facebook users can upload videos and share the videos with their friends. Also Facebook users are able to upload videos from YouTube or other online video sharing web sites in Facebook as well. Therefore, the users can link online video sharing web



sites to online social networks. One of the most popular online social networks is YouTube. So in this section, the related works about YouTube and its online video sharing network characteristics are presented.

In YouTube, each user can add friends and watch their friends' uploaded video files. Cheng *et al.* [13, 27] did some research on some YouTube features. The small average number of friends and the small median of friends numbers in YouTube user network presents that the social network among users is not as strong as social network among video files existing in YouTube.

User level characterization in YouTube presents that the mean and median numbers of the transactions per users are 152 and 51. As a result of various users' behaviour, the user's transaction number (the number of access to the video files) is very different. Similar to the transaction numbers per user, the amount of transferred data per user varies because of the large size of video clips. The result of transferred content type presents that there are four different kinds of content types: images which are most transferred data followed by text, applications and videos. Although video content is the least transferred content type, most of the transferred bytes is belonged to the video content.

Maia *et al.* [28] did the research on classifying user behaviour of YouTube. By applying BFS algorithm, K-mean clustering and collecting some information from users' profile such as the number of views and the number of uploaded videos, the following five groups are identified. The first identified user group is a small community member: this group consists of users who have interaction out of network as well such as friends, colleagues, or families members as a result of a high clustering coefficient and small in-degree and out-degree. The high clustering

coefficient means that they have strong friendship connection in their network; also they do not have many friendship links with users out of their network. The small in-degree and out-degree means that they upload, watch, or subscribe a few videos probably being that it is their friends' videos. The next group is the content producer which is 23% of all users, and they have lots of uploaded videos. The third group is content consumers, which prefer to watch lots of videos instead of uploading videos. The next group is producer and consumers, which are the largest group, about of 48% of all YouTube users. They have both characteristics of producer and consumers. In other words, they have a large number of uploaded and viewed videos. The last group is other users, which cannot be clear identified as a result of the low value of every feature.

The next step is recognizing the main characteristics of user behaviour. Timing is not an important feature in identifying user behaviour. It means that the newer users have almost the same behaviour as the older users. The result also presents that social users' interaction has the critical role in identifying users' behaviour. This result is not only for YouTube, but also is applicable for all other online social networks.

Gill *et al.* [11, 12] characterized user sessions on YouTube and then compared it to the traditional web workloads. Session duration, inter-transaction times, and the types of the content transferred by user session are considered as characteristics of the user session. There is a limitation of 100 MB for file size in YouTube. The small fraction of videos with more than 100 MB size shows that file size limit is not quite strict in YouTube. The shorter videos tend to be more popular. Also the result shows that the most viewed videos of a week and a month has less than one week or one month age. The result of rating of videos shows that more than 80% of the time, the average rating is more than 3. The short term popular videos are mostly belonging to the entertainment, sports and news categories. Moreover, the long term popular videos belong to

the comedy, entertainment and music categories. The least popular categories are Auto and vehicles, Howto and DIY (Do It Yourself), pets and animals and travel and places.

Some YouTube videos' characteristics are analyzed such as the length, access pattern, their growth trend and active life span by Cheng *et al.* [13, 27]. Then some of these statistics are compared with traditional web servers. The comparison between age of the video files and the number of the views presents that the older videos have more chances to become popular. However, there are some young popular videos and old unpopular videos.

As it is mentioned, some videos become popular in the short amount of time; whereas some others never were accessed after a while. The result shows that the power law distribution can fit the growth trend better than linear. Some video popularity grows more and more slowly until it stops, which is called active life span for a video. Most of the videos have a short active life span. In YouTube, the number of comments is less than the rating, and both of them are less than the number of views.

#### **2.2.1.3 Other Online Social Networks Workload Characterizations**

Four popular online social networks have been analyzed: Flickr, LiveJournal, YouTube and Orkut by Mislove *et al.* [14]. Their network consists of users who have online social activity, so the users who do not have online activity (isolated users) are not considered in the graph. To analyze the existence of short path length in the graph the large weakly connected component (WCC) of the corresponding graph is considered in the research instead of the whole graph. They characterized different features of these networks. The first feature is the number of in-degree and out-degree of nodes that presents that for all of four online social networks: Flickr, Orkut, LiveJournal, and YouTube, out-degree and in-degree node distributions follow the power law

distribution. Grouping is one of the features that make the online social networks more interesting. It means that the users with the same interests make a group in the network. The network users who are member in the same group are not necessarily connected to each other. Their result presents that the group size in these four networks follows the power law distribution.

Mitra *et al.* [29] characterized four different video sharing web sites: Dailymotion, Yahoo Video, Veoh, and Metacafe. Metacafe. The cumulative distribution of the total views popularity and the cumulative distribution of the viewing rate popularity show that both of them are following the Pareto principle. It means that only 20% of the most popular video clips are sufficient to account for 80% or more of the views.

For Yahoo Video and Metacafe, the CCDF plot of the total views popularity distribution fits the power law distribution. However, for Dailymotion and Veoh, both power law exponential cut off (exponent between 1.4 and 2.3) and the log normal distribution fit the model. The result of two weeks snapshots presents that average viewing rate over the certain time period is more Zipf-like than average viewing rate since upload, and average viewing rate since the upload is more Zipf-like than total views popularity.

Kang *et al.* [7] worked on internet video sharing site workload, Yahoo Video. The static properties show that the mean video clip duration is 283.46 seconds and the median video clip duration is 159 seconds. In addition, the result shows that Zipf distribution with an exponential cutoff fits file popularity pattern. The single video server is modeled with First Come First Served model (FCFS). In this model capacity is defined as the number of the stream delivered by the server in the same time without losing any quality. Two scheduling schemes are chosen: one

of them is random dispatching, which sends the request to a random server. The other scheme is Least Workload Left (LWL) which sends the request to a server with the least remaining workload. Two QoS (Quality of Service) are considered as Service Level Agreement (SLA): quality of video clips and waiting time of the video stream in a queue. It is assumed that there is enough bandwidth to send the video stream to the users; thus, the quality of video stream is guaranteed. Collecting one week measured data on random and LWL schemes shows that 69.9% of servers could be saved by using LWL schemes in comparison to random dispatching scheme.

### **2.2.2 Study the Small World Characteristic in Online Social Networks**

As described by Cheng *et al.* [13, 27], small world phenomenon is one the most important features of user network graph in social networking sites. Therefore, the existence of small world in YouTube network is presented in section 2.2.3.1, and existence of small world in other online social networks is presented in section 2.2.3.2. In the best of our knowledge there is no similar work on examining the small world characteristic for Facebook users.

#### **2.2.2.1 The Small World in YouTube**

As it is mentioned by Cheng *et al.* [13, 27], the existence of small world characteristic is one of the most interesting characteristic of online social networks. As it is mentioned before, the network should have two following conditions to have the small world characteristic: first it should have a small average short path length and secondly it should have a high clustering coefficient.

The definition of user networks would not be the same for each site in different social networks. For example, Cheng *et al.* [13, 27] defined each video file as a node in the YouTube network graph. In this graph, the list of related video files of a node is defined as the neighbours

of the node. They presented that this graph has the high clustering coefficient and the small short path length characteristics. Therefore, YouTube's related videos network has small world characteristics. This means any two video files are reachable by the small number of links.

#### **2.2.2.2 The Small World in Other Online Social Networks**

As it is presented in the last section Mislove *et al.* [14] did the research on four popular online social networks: Flickr, LiveJournal, YouTube and Orkut. They calculated the average short path length and the average clustering coefficient of their corresponding network graphs. The results present that these networks have shorter average path length in compare with the web graph, and all of these four networks have a higher average clustering coefficient than the web. Due to the existence of small world phenomenon in the web graph in which the web page is considered as a node and the web page links to the other pages considered as the edge, all of these four online social networks have the small world characteristic as well.

## Chapter 3

### 3. Methodology and Measurement

As mentioned in the introduction of this thesis, we want to model Facebook user network. So that user networks in Facebook can be analyzed. We have published paper that shows the result of characterization of two user networks in Facebook [30]. One of these networks is related to the user profiles and their friends. By using this network, two users in Facebook can link to each others. The other network is related to the web page fans. As mentioned before, there are many web pages in Facebook each including their own fan members. By using these networks, news related to these web pages can broadcast to any other users of Facebook even if they are not part of web pages network.

As it appears in the related works section, there are some researches on Facebook applications. However, there are few researches on Facebook web pages, which have great effects on spreading news of marketing companies, political events, and etc. As it is shown in table 3.1, two data sets are collected in this work.

Table 3.1 Collected data sets

Data Set number	Source	Period	Collected data
1	Facebook unblocked user profiles	Oct15,09- May 15,10	20,019 Users and their friends
2	Facebook web pages	July01,09- Nov 8,09	1, 048 fans for 8 web pages

For modelling the data first we generated their CDF plots. Then we used visual observation for the empirical distribution and candidate models that are suggested in previous works followed by performing goodness of fit (K-S test) as suggested in [31]. We validated the achieved curve fitting results by running ExpertFit [32] software.

As it is mentioned by Cheng *et al.* [13, 27], one of the most important features of online social networks is small world characteristic. In this thesis, the concept of the small world phenomenon will be further examined.

The rest of this chapter is organized as follows: user network characterization is presented in section 3.1, followed by web page fans characterization in 3.2. The examination of existence of the small world phenomenon is presented in section 3.3.

### **3.1 User Network Analysis**

Similar to other works, the Facebook user network will be modeled as a graph, so that in the graph, Facebook users are represented as the nodes. The nodes connect to other nodes by their friendship link in Facebook. Thus, in our research, the number of friends represents the node degree in the graph. In Facebook network graph, links are undirected; in other words, if user A is friend with user B, then user B is friend with user A as well. So in the Facebook graph, the in-degree of a node is same as its out-degree. For this part, a realistic way is needed to gather data from Facebook, which helps us in gathering each user's names and number of the user's friend's.

The reasonable way to gather data from Facebook is similar to the method that is presented by Nazir *et al.* [3], is generating several user profiles in Facebook and joining them to



different network cities or network countries. The first reason of doing that is simplicity; for example, for joining university networks, university email address is needed. Second, many users allow for users from the common regional networks as theirs to access their profile information.

For collecting the first data set, twenty fake user profiles are generated, and they were joined different countries and cities networks such as: Sweden, India, China, Japan, and New York. To crawl the data, one user from the networks is chosen randomly similar to the method in [5]. Then in each step the list of the user's friends that is not visited yet and their number of friends is added to our data set. This algorithm is continued until enough data for our measurement study was gathered. This algorithm is called BFS. This kind of data gathering has limitations as well. First of all, some users' profiles are blocked even for users from the same network. Secondly, some of the Facebook users do not join any networks, so they are not reachable through the use of this method. In our research, the users who are unreachable by other users are not considered in this network. In other words, isolated users are not considered in our graph. Also the users with blocked profile are not considered in our network.

The data set consists of numbers of users' friends and list of their friends. Our data set is much smaller than the whole Facebook network. Therefore, to increase the reliability of our data set, two smaller data sets are also collected. The characteristics of those two data sets are similar to our data set. So we did not include them in the graphs of the next section.

In our work, the number of a user's friends is a node degree. To characterize this network in Facebook, we try to find the node degree distribution. Figure 3.1 presents the Cumulative Distribution (CDF) of node degree examining the Lognormal distribution and the exponential

distribution. The statistics of Facebook's node degree (embedded objects) are shown in Table 3.2.

The Kolmogorov-Smirnov test (K-S test) is applied in order to find the suitable distribution for the Facebook node degree distribution. In the K-S test the distribution with the lower value is the best match. As table 3.3 and visual comparison suggest the node degree distribution of users in Facebook follows a Lognormal distribution.

Table 3.2 Statistics of the node degree of Facebook

Characteristic	The node degree of Facebook
Mean	405.9
Median	241
Standard Deviation	616.48
Maximum	5,002
Sample Size	20,019
CV	1.51
Model	Lognormal a= 5.41 b=1.19

Table 3.3 Kolmogorov-Smirnov Test

Model	K-S Statistics
Lognormal distribution	0.05497
Exponential distribution	0.11620

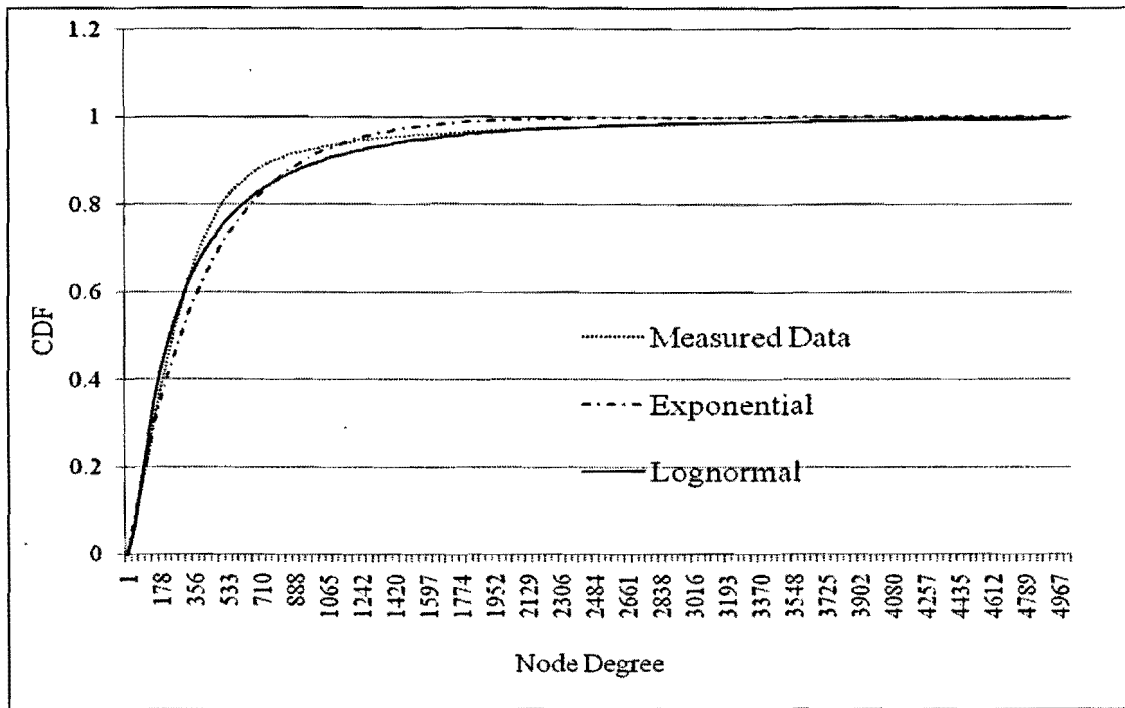


Figure 3.1 The node degree distribution of Facebook

### 3.2 Facebook Web Page Fans

As mentioned in the introduction, several web pages in Facebook are developed for various TV shows, political faces, celebrities, public places, and marketing companies. Facebook users can become a fan of these web pages and the news of these web pages can be broadcasted widely. The number of web page fans increases each day; therefore, the news of the web pages broadcasts spreads with more speed each day. This frequent change shows the dynamic nature of the number of Facebook users, which we decided to study in this work. To characterize web page fans, the increase in the number of web page fans was necessary. For collecting the second data set several web pages that are chosen from different categories such as Starbucks and Timhortons from marketing companies, CNN and MTV from TV channel, Oprah show and

Friends from TV shows category, Facebook web page from entertainment and two other random famous web pages.

Our data set consists of the number of web page fans, date, and web page name for each day during a 131 day period, from July 1, 2009 until November 8, 2009. Similar to the Facebook, the web page usage decreases over weekends and holidays.

The distribution of the increase in the number of web page fans in Facebook shows that the distribution can be modeled by the Weibull distribution. Figure 3.2, 3.3, 3.4, 3.5 present the Cumulative Distribution (CDF) of the increase in the number of fans of Starbucks, CNN TV channel and MTV channel and Facebook web page. To find the model for increase in the number of fans in the Facebook web pages, we examining the Weibull and Gamma distributions.

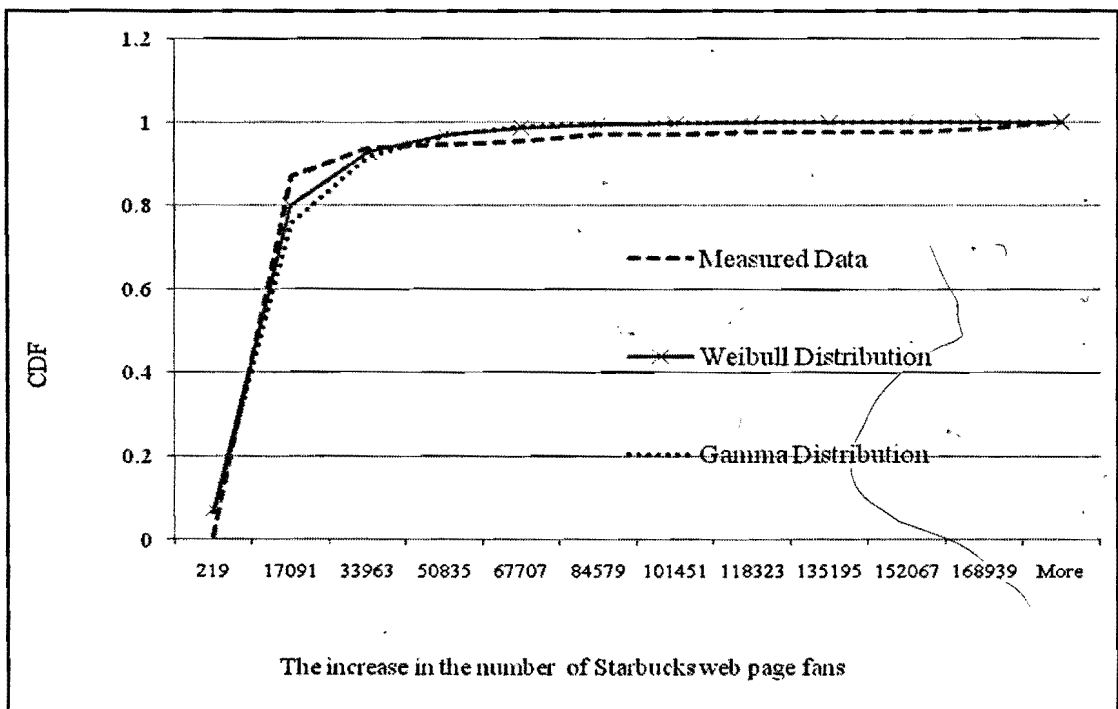


Figure 3.2 The distribution of the increase in the number of Starbucks web page fans

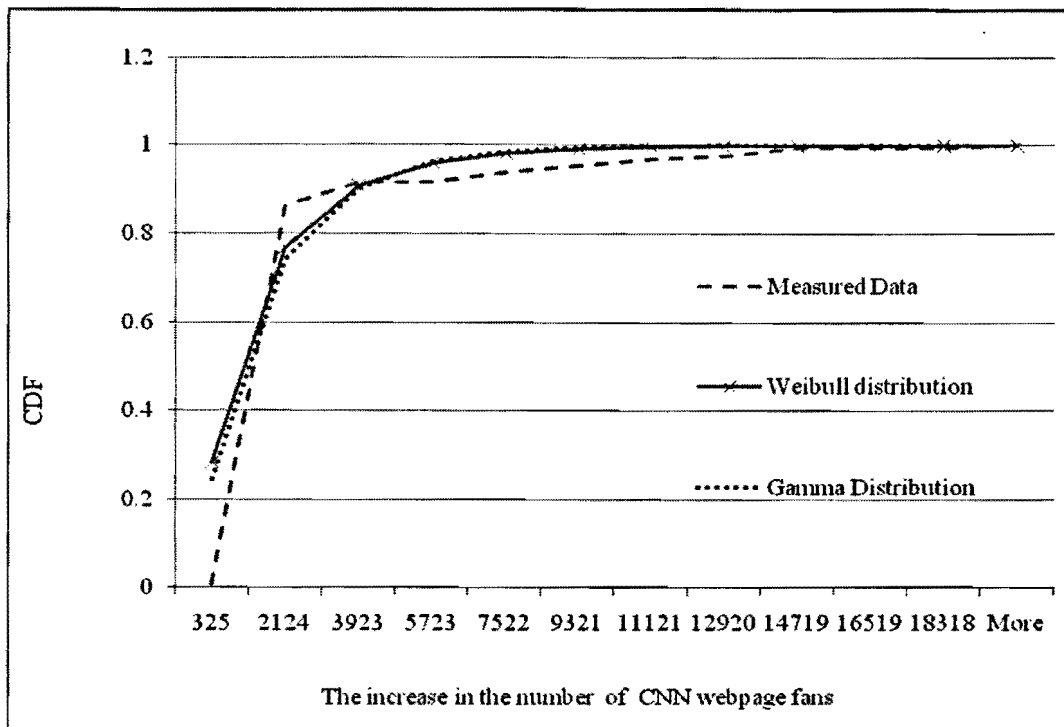


Figure 3.3 The distribution of the increase in the number of CNN web page fans

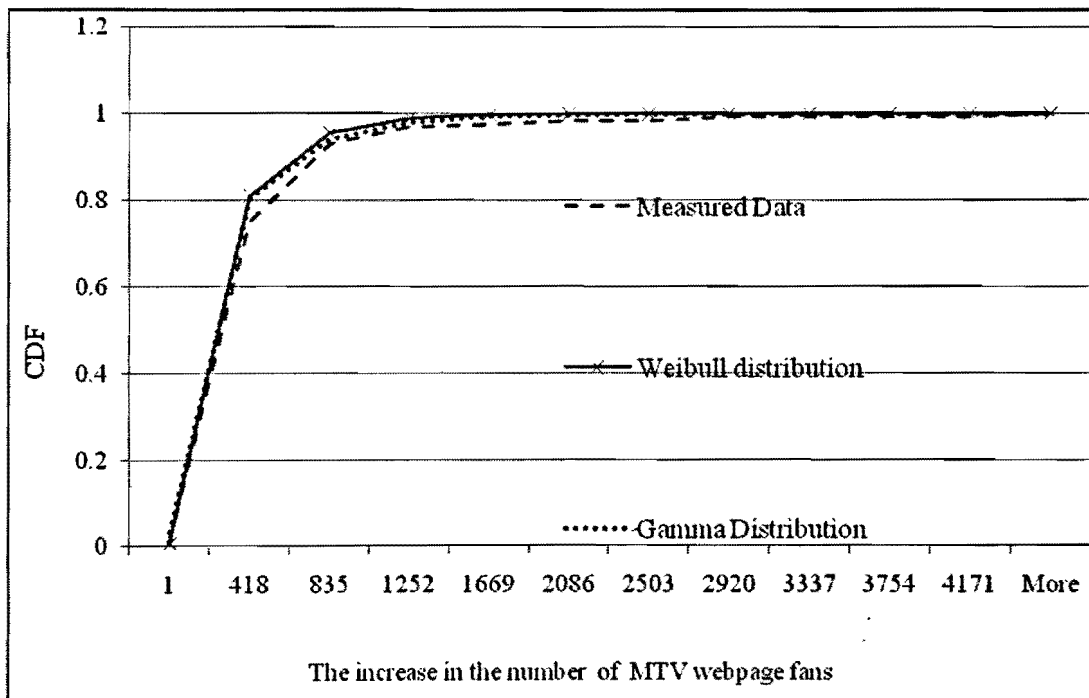


Figure 3.4 The distribution of the increase in the number of MTV web page fans

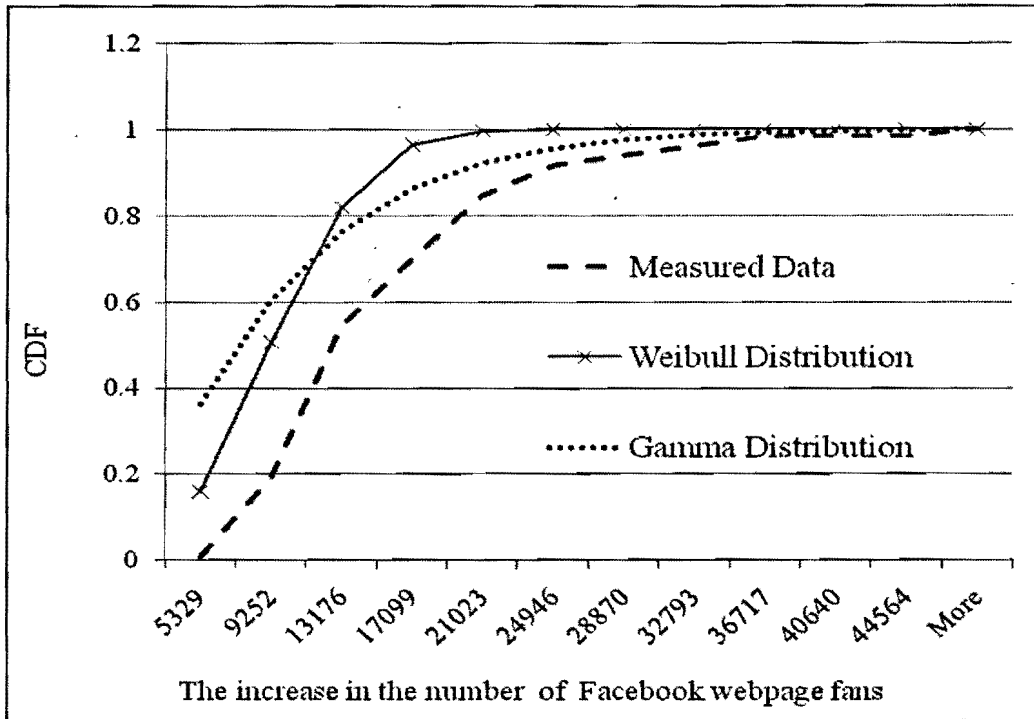


Figure 3.5 The distribution of the increase in the number of Facebook web page fans

The Kolmogorov-Smirnov test (K-S test) is also applied to find the suitable distribution for modelling the increase in the number of the fans for these four Facebook web pages fans. As previously mentioned, in K-S test the distribution with lower value is the better match. As table 3.5 and visual comparison suggest the distribution fits the Weibull distribution.

The statistics of increase in the number of these three Facebook web page fans are presented in Table 3.4.

Figure 3.6, 3.7, 3.8, and 3.9 present Cumulative Distribution (CDF) of increase in the number of Timhortons, Friends TV show, Oprah TV show, and a political person web page fans examining the Weibull and Gamma distribution as the another good candidate. The statistics of the increase in the number of these four Facebook web page fans are presented in table 3.6.

Table 3.7 presents the Kolmogorov-Smirnov test (K-S test) to find the suitable distribution for the increase in the number of these four Facebook web page fans. As the table 3.7 and visual comparison suggest the best fit to model our collected data is the Weibull distribution, which agrees with web page fans.

Table 3.4 Summary statistics of the increase in the number of fans in four web pages fans

Characteristic	Starbucks	CNN TV Channel	MTV	Facebook
Mean	12,327.67	1,895.3	347.37	14,667.43
Median	3,833	953	234	12,027
Standard Deviation	29,122.36	2,984.93	521.52	7,617.54
Maximum	185,811	20,118	4,588	48,488
Minimum	219	325	1	5,329
Sample Size	130	130	129	130
CV	2.36	1.57	1.5	0.51
Model	Weibull a=0.7 b=8714.76	Weibull a=0.78 b=1311.25	Weibull a=0.91 b=239.9	Weibull a=2.56167 b= 10,637.78

Table 3.5 Kolmogorov-Smirnov Test

Model	K-S Statistics (Starbucks)	K-S Statistics (CNN TV Channel)	K-S Statistics (MTV)	K-S Statistics (Facebook)
Weibull Distribution	0.69005	0.23545	0.05497	0.3128
Gamma Distribution	0.109935	0.23793	0.5573	0.4108

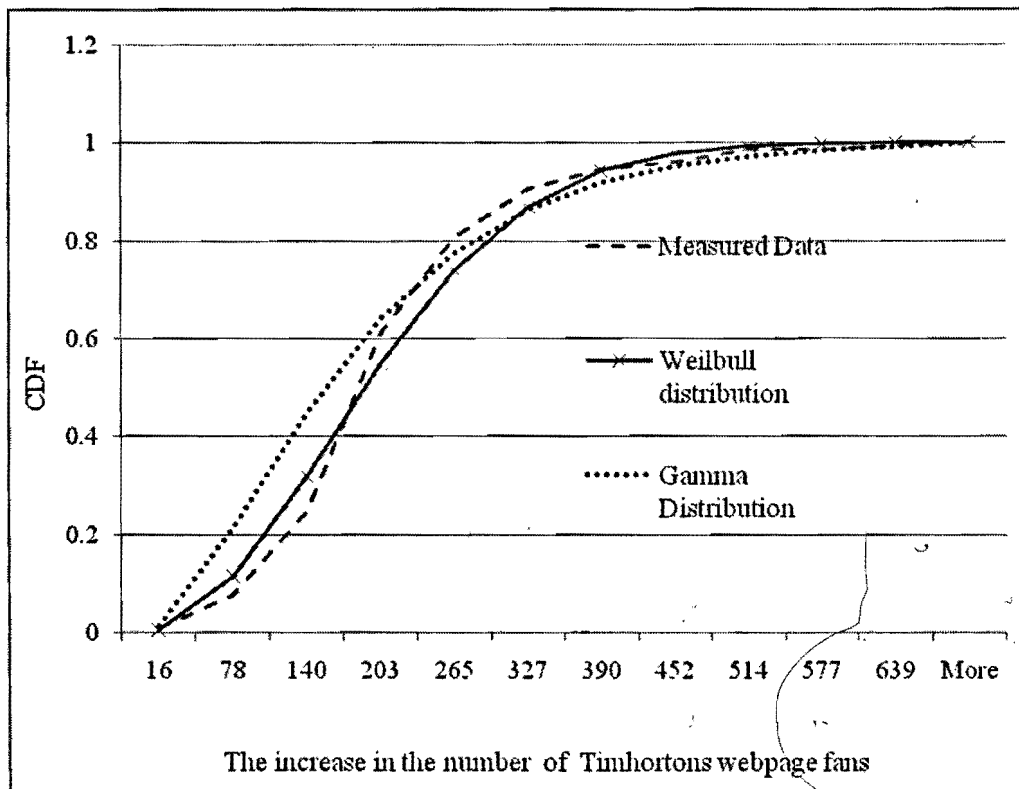


Figure 3.6 The distribution of the increase in the number of Timhortons web page fans



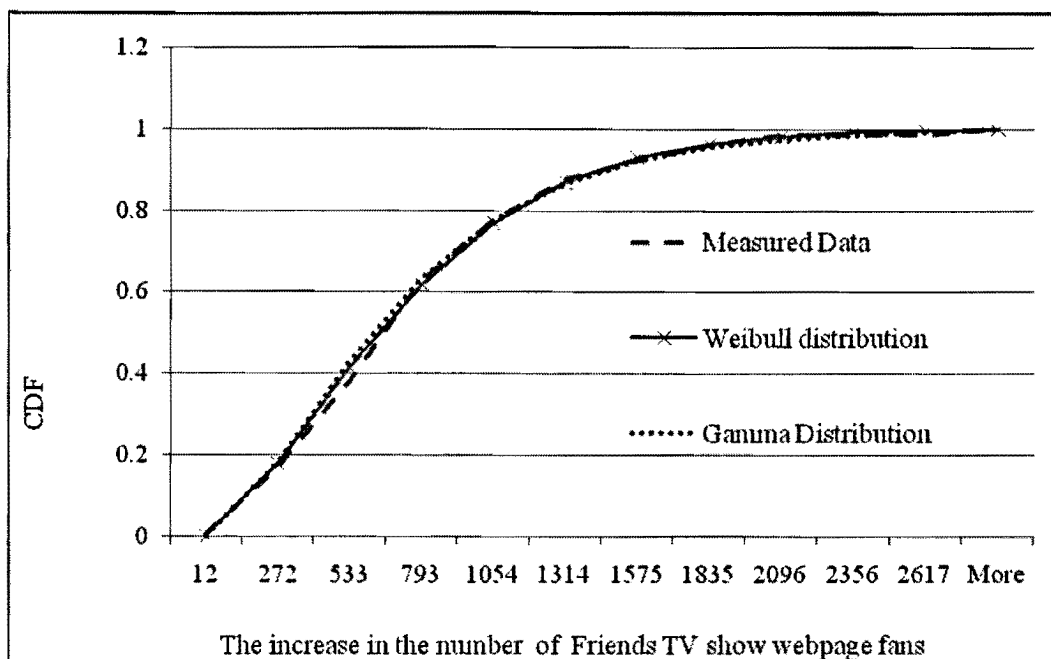


Figure 3.7 The distribution of the increase in the number of Friends TV show web page fans

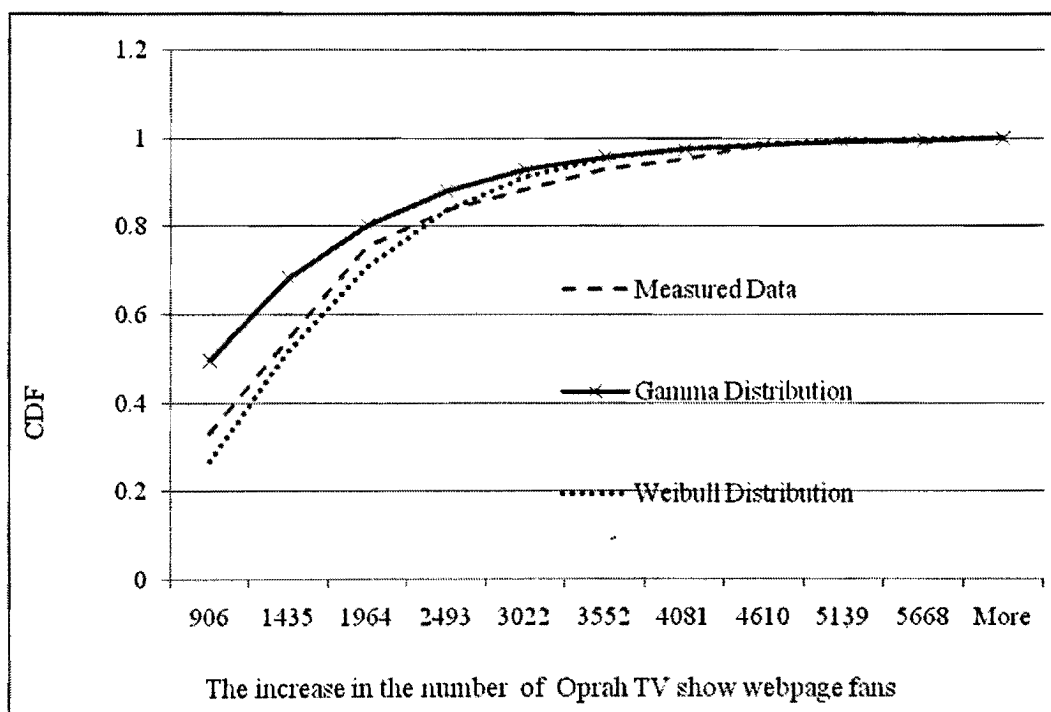


Figure 3.8 The distribution of the increase in the number of Oprah TV show web page fans

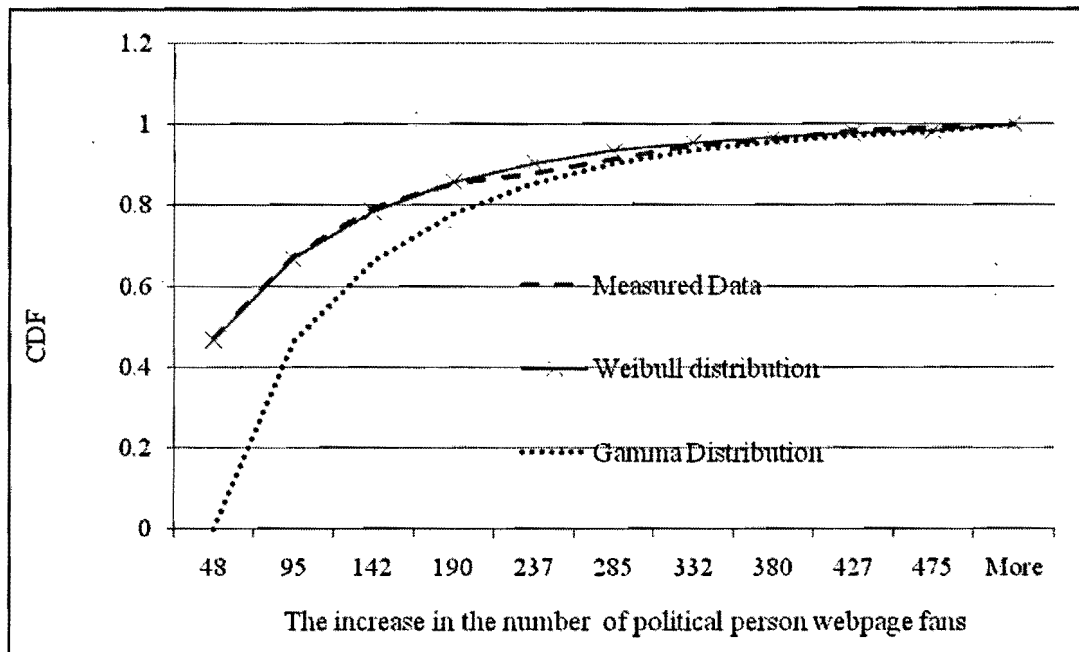


Figure 3.9 The distribution of the increase in the number of a political person web page fans

Table 3.6 Summary statistics of the increase in the number of fans in four web pages fans

Characteristic	Timhortons	Friends TV show	Oprah TV show	a political person
Mean	201.38	747.54	1,594.62	92.5
Median	184	635	1,342.5	50.5
Standard Deviation	108.05	509.09	1,098.51	108.23
Maximum	702	2,878	6,198	523
Minimum	16	12	377	1
Sample Size	130	131	130	130
CV	0.53	0.68	0.68	1.17
Model	Weibull a=1.95 b=227.8	Weibull a=1.47 b=813	Weibull a=1.11 b=1269.09	Weibull a=0.82 b=84.23

Table 3.7 Kolmogorov-Smirnov Test

Model	K-S Statistics (Timhortons)	K-S Statistics (Friends TV Show)	K-S Statistics (Oprah TV Show)	K-S Statistics (a political person)
Weibull Distribution	0.0721	0.03	0.063	0.0265
Gamma Distribution	0.201	0.0488	0.1658	0.4692

### 3.3 Small World Characteristic in Facebook

In this part, we want to examine the existence of small world phenomenon in Facebook user network. A graph has small world characteristic, if it has two conditions: firstly, the graph should have a high average clustering coefficient, secondly, it should have a small average short path length.

We have submitted journal paper examining the existence of small world characteristic in Facebook [33]. Same as section 3.1, we use a graph to model Facebook user network, while defining each user as a node and their friends link as the edge of the graph. In this section, we use BFS algorithm to collect our data set. Same as the previous section, a user is selected randomly from the Facebook user network. Then by using BFS algorithm and twenty fake user accounts that are created in the last section, the data set can be collected. The data set consists of the user name and the list of its friends. The BFS algorithm is continued until enough data has been gathered. Therefore, data set 1 is increased to 24,799 Facebook users:

Section 3.3.1 presents our methodology to calculate the average clustering coefficient of the graph that followed by our proposed methodology to find the average short path length in the graph that is explained in section 3.3.2.

### 3.3.1 Average Clustering Coefficient

As mentioned before, the clustering coefficient of a node is defined as the number of all the existing links between node's neighbours divided by the number of all possible edges between the node's neighbours. The clustering coefficient of a graph is defined as the average of clustering coefficient of all graph nodes. The following algorithm is used in our thesis to calculate the average clustering coefficient of our collected data for Facebook user network:

---

**Algorithm 3.1** Calculation of the average clustering coefficient in a graph  $G(n)$

---

```

read n as the number of nodes
sum  $\leftarrow$  0
for each  $i \leq n$  node do
    z  $\leftarrow$  i's node degree
    k  $\leftarrow$  number of edges between node i's neighbours
    clustering_coefficient  $\leftarrow$  calculate the clustering coefficient by using formula 1.7
    sum  $\leftarrow$  sum+ clustering_coefficient
end for
average_clustering_coefficient  $\leftarrow$  sum/n

```

---

To explain the process of computing the average clustering coefficient of the graph, first we define term "n" as the number of nodes in the graph.

To calculate the average clustering coefficient of the graph, at first the average clustering coefficient of each node of the graph should be calculated. To calculate the clustering coefficient of each node, the first step in the algorithm is to find the number of a node degree for the first

node. To calculate the node degree, we should go through our data set and find all the neighbours of the node. In the next step, the number of existing edges between the node's neighbours is calculated. At the end, by using formula 1.7, the clustering coefficient of the node is calculated. As mentioned before, to examine the existence of small world characteristic, the average clustering coefficient of the graph is needed. So this algorithm should be continued until the clustering coefficient of all nodes is calculated. At the end of the algorithm, the average clustering coefficient of the graph is computed.

### 3.3.2 Average Short path Length

As it is mentioned in section 2.1.5.1.2, in a mathematical graph, the path length is defined as the average of the minimized number of hops or steps between any two nodes in the graph. There are several algorithms to find the average short path length of the graph. Due to the size of our graph, most of these algorithms are found to be very time consuming and can occupy a large volume of memory space. In this research, breath first search algorithm is applied to find the average short path length of the graph which is presented by Cormen *et al.* [34].

In this algorithm, two corresponding arrays are applied in order to calculate the average short path length of each node. To calculate the average short path length of each node, the nodes' number for all nodes is saved in the first array, and the shortest path length between a specified node and the corresponding node from the first array is saved in the second array. In the first step of the algorithm, the short path length between the node and itself is 0. Therefore, the number of specified node is saved in the first array, and zero which is the shortest path length between the node and itself is saved in the corresponding element of the second array. In the next step, the node's neighbours are saved in the first array, and the short path length between

the node and its neighbours is  $0+1=1$  which is saved in the corresponding elements of the second array. In the next step, the neighbours of the nodes that are saved in the first array in the previous step are identified, and if the shortest path length between the node and those nodes are not calculated yet, those nodes are added to the first array. The short path length for those nodes is equal to the short path length of their neighbour's with the smallest short path length plus one. This process is continued until the short path length between the node and all other nodes in the graph are calculated. Then the average short path length of that node is computed. As previously mentioned, the average short path length should be calculated for all graph nodes. Therefore, this algorithm is continued until the average short path length of all the nodes in the graph is computed. At the end of the algorithm, the average short path length of the graph is calculated.

### 3.3.3 Discussion of the Existence of Small World Characteristic

In the last two sections, the average clustering coefficient and the average short path length algorithms were presented. According to Watts *et al.* [24] the small world network has smaller average short path length and higher average clustering coefficient than the Erdos-Renyi [35, 36] random graph. Therefore, to examine the existence of small world characteristic, the Erdos-Renyi random graph with the same number of nodes and links should be generated. By applying the algorithm 3.2, the Erdos-Renyi random graph is generated with the same number of nodes and links similar to the graph which is denoted in the previous section.

In Erdos-Renyi random graph, the edge between any two nodes has equal and the independent probability.

---

**Algorithm 3.2** Generating Erdos-Renyi Random Graph  $E(n,p)$  from [35, 36]

---

```
read n as the number of nodes
read p as the probability of each edge
// p is defined as the number of existing edges in the graph divided by the number of all possible edges in
the graph//
G ← Generate a upper triangular part random matrix (n*n)
for each  $i \leq n^2/2$  elements of upper triangular part of matrix G do
    if  $G(i,j) < p$ 
         $G(i,j) \leftarrow 1$ 
    else
         $G(i,j) \leftarrow 0$ 
    end if
end for
//erdos is the adjacency matrix of the generated Erdos-Renyi random graph//
for  $i \leq n$  do
    for  $j \leq n$  do
         $erdos(i,j) = G(i,j) + G(j,i)$ 
    end for
end for
```

---

Erdos-Renyi random graph algorithm has two inputs: “n” as the number of nodes, and “p” as the probability of each edge. In the first step, a random upper triangular matrix with elements between 0 and 1 is generated. Then each element of the matrix is compared with “p”, and then if the element value is less than “p”, the element value will be changed to 1; otherwise, it will be changed to zero. At the end of the algorithm 2 the adjacency matrix of the Erdos-Renyi random graph is created.

The average clustering coefficient and the average short path length of the Facebook user network graph are computed by the algorithms that are presented in the previous sections. Then by applying algorithm 3.2, the Erdos-Renyi random graph with the same number of nodes and links as our Facebook user network graph is generated. The average short path length and the average clustering coefficient of the generated Erdos-Renyi random graph are computed as well

by applying the algorithm 3.1 and the average short path length algorithm. As it is presented in table 3.8, the average short path length and average clustering coefficient of the Erdos-Renyi random graph and our graph are compared.

Table 3.8 The comparison of between Facebook user network and Erdos-Renyi random graph.

Characteristics	Number of nodes	Erdos-Renyi random graph	Facebook user network graph
average clustering coefficient	24,799	5.37 E-5	5.67E-2
average short path length	24,799	8.60	5.74

As shown in table 6, our Facebook user network graph has a higher average clustering coefficient and smaller average short path length than Erdos-Renyi random graph. Therefore, as it is mentioned in [18], Facebook user network formed in our collected data has small world characteristic.

### 3.4 Conclusions

In this chapter, two data sets are collected from Facebook networks. The first data sets consists the Facebook user names and lists of their friends. The result presents that the node degree distribution of user network in Facebook fits the Lognormal distribution. The second data sets consists of the number of increase in the number of fans belonging for several web pages in Facebook such as Starbucks, Timhortons, MTV, Friends TV show, Oprah TV show, and other popular web pages in Facebook during 131 day period. The result presents that the number of increase in the number of different web pages fans in Facebook follows the Weibull distribution.



Furthermore, the algorithms for calculating the average clustering coefficient, the average short path length of the graph and generating the Erdos-Renyi random graph are presented in this chapter. The comparison of the average clustering coefficient and the average short path length between the Facebook user network in our collected data graph and the Erdos-Renyi random graph with the same number of nodes and links presents that the Facebook user network has smaller average short path length and higher average clustering coefficient. Therefore, at the end of this chapter, we concluded that the Facebook user network graph has a small world characteristic. In other words, any two users in Facebook network can link to each other by small number of hops.

## Chapter 4

### 4. Synthetic Workload Simulator

As mentioned before, the contribution of this research is building a synthetic workload simulator of the Facebook network. To pursue this goal, the graph with the same characteristics which are presented in the previous sections is generated. Several random network graph generators with different properties are presented in [37, 38, 39]; however, none of these algorithms can generate the network graph with the specific node degree distribution and the small world characteristic. The method for generating Facebook user network graph is presented in section 4.1, and section 4.2 presents the experiment. Section 4.3 presents the methods for generating web page fans graph and validating the software, Finally section 4.4 presents the conclusions of this chapter.

#### 4.1 Generating Facebook User Network Graph

In this section, a network graph with the same attribute as the Facebook user network graph is generated. In other words, a small world random graph with the Lognormal node degree distribution should be generated. The first step is generating the appropriate node degree distribution that is presented in section 4.1.1. Section 4.1.2 presents the process of generating the graph. Finally, generating the small world network graph is presented in section 4.1.3.

##### 4.1.1 Generate the Node Degree Distribution

As mentioned before, in the first step, we should generate the random graph with the same node degree distribution (as it is concluded in section 3.1 the node degree distribution of

Facebook user network fits the Lognormal distribution). Our algorithm is similar to the algorithm that is proposed by Newman *et al.* [40]. At first, we developed a random number generator program to generate “n” random numbers between 0 and n-1 that fits the Lognormal distribution. These numbers indicate the list of the node degree sequence of the new graph. So there is limitation of n-1 on the created number. The reason for that is in the graph each node has the chance to be connected to maximum n-1 nodes.

In a graph, number of all in-degrees is equal to the number of all out-degrees, so the summation of the node degrees (both in-degree and out-degree) cannot be odd number. Thus, if the summation of all the generated numbers is odd, then the generated numbers are not qualified for the node degree distribution. In that case, one of the numbers is selected randomly, and then subtracted by 1. Therefore, at the end of this step, the summation of the generated numbers is even, so they are qualified for the required node degree distribution.

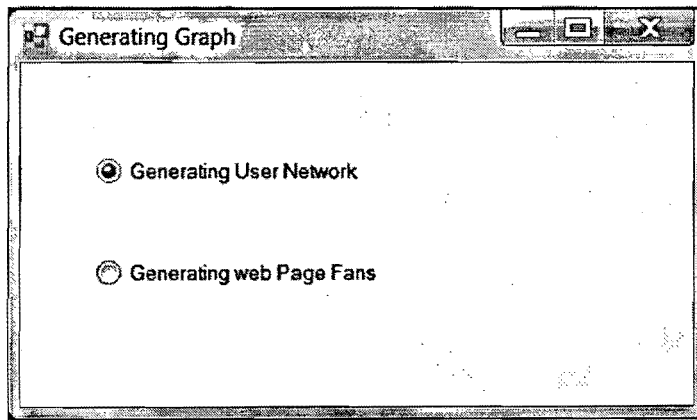


Figure 4.1 Generating Graph form

Figure 4.1 and 4.2 are the user interfaces of our developed software to generate Facebook graphs. User network or web page fans graphs can be generated by selecting the related option from the form shown in figure 4.1, the default values for the related distribution can be selected by the form shown in figure 4.2. Note that the big advantage of the synthetic workload simulator

is being parametric. It means that our program can accept any values from the user and generates the resulted graphs that can be used in variety of scenarios with different needs.

Generating User Network Graph

Lognormal default value

Scale 214.72

Shape 1.19

Size 100

Create Data

Created Data (Node degree)

2  
3  
13  
2  
6  
5  
6  
1  
12  
3

To simulate the Facebook user network graph by generated datapress NEXT

NEXT>>

Figure 4.2 Generating User Network Graph form

#### 4.1.2 Generating the Graph

The next step of generating the network graph is to generate the graph with the same node degree distribution and random connection of nodes to each other. To generate the graph, two corresponding arrays are applied for which the first array consists of the node sequence, and the second array consists of the node degree number. In each step of the algorithm, two random non zero node degrees are selected from the second array. If the corresponding nodes from the first array are not connected to each other yet, we connect these two nodes in our graph. Then their node degree is subtracted by one in the second array. This algorithm is continued until the

entire node degrees become zero in the second array. By the end of this algorithm, the new graph with the Lognormal distribution of node degree is generated.

#### **4.1.3 Generating the Small World Network Graph**

As mentioned before, the small world network graph has two major characteristics. Primarily, the generated graph should have a higher clustering coefficient than the Erdos-Renyi random graph that has the same number of nodes and links as the generated graph. Secondly, it should have smaller average short path length than the Erdos-Renyi random graph.

Therefore, the first step is generating the Erdos-Renyi random graph. So, by applying algorithm 3.2, the Erdos-Renyi random graph with the same number of nodes and the links same as the generated graph is created. In the next step, the small world network graph should be generated. Section 4.1.3.1 presents the algorithm to optimize the average clustering coefficient of the graph in order to generate the network graph with a higher average clustering coefficient than the generated Erdos-Renyi random graph. Section 4.1.3.2 presents the second required condition of small world network which is generating the network graph with a smaller average short path length than the Erdos-Renyi random graph.

##### **4.1.3.1 Optimizing the Average Clustering Coefficient Value**

At first, by applying algorithm 3.1, the average clustering coefficient of the generated graph and the Erdos-Renyi random graph that is produced in the previous section are calculated. The average clustering coefficient of the generated graph is compared with the average clustering coefficient of the Erdos-Renyi random graph. Then, if the generated graph has a higher average clustering coefficient than the Erdos-Renyi random graph, then the generated random graph has the first condition of small world characteristic.

If the graph has smaller average clustering coefficient than the Erdos-Renyi random graph, then the average clustering coefficient of the generated graph should be tuned. Algorithm 4.1 which is similar to the presented algorithm by Guo *et al.* [41] will be applied to increase the average clustering coefficient of our generated graph as much as it becomes larger than the average clustering coefficient of the generated Erdos-Renyi random graph.

---

**Algorithm 4.1** Optimize the average clustering coefficient Cluster( $C_E, C_G$ , counter)

---

read  $C_E$  as the average clustering coefficient of Erdos-Renyi random graph  
read  $C_G$  as the average clustering coefficient of generated graph

while  $C_G < C_E$  and counter  $> 0$  do

$Q \leftarrow$  Select a random node in the graph

$(n1, n2) \leftarrow$  Select two random neighbours of node  $Q$

$m1 \leftarrow$  Select a random neighbour of  $n1$

$m2 \leftarrow$  Select a random neighbour of  $n2$

  // conditions:

  1) The edges between  $(n1, n2)$  and  $(m1, m2)$  do not exist

  2)  $(n1, n2, m1, m2)$  are not the same node//

  if  $(n1, n2, m1, m2)$  meeting the *conditions* do

    delete link between  $n1$  and  $m1$

    delete link between  $n2$  and  $m2$

    add link between  $n1$  and  $n2$

    add link between  $m1$  and  $m2$

$c2 \leftarrow$  the average clustering coefficient of the new graph

    if  $c2 < C_G$

      Change the whole graph to the initial situation

    else

$C_G \leftarrow c2$

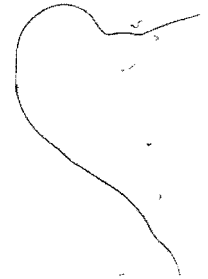
    end if

  end if

  counter  $\leftarrow$  counter-1

end while

---



In this algorithm, we try to optimize the average clustering coefficient by swapping the edges between the nodes. In the first step of the algorithm a random node from the generated graph is chosen. Then two of the node's neighbours are selected ( $n_1, n_2$ ). One of the neighbours of  $n_1$  and one of the neighbours of  $n_2$  are chosen randomly ( $m_1, m_2$ ). The selected nodes ( $n_1, n_2, m_1, m_2$ ) should have two conditions to go through the next step: first,  $n_1$  should not be connected with  $n_2$  and also  $m_1$  should not be connected with  $m_2$ . Second, these four nodes should not be the same. If the nodes have these conditions the links between  $n_1$  and  $m_1$  and between  $n_2$  and  $m_2$  are deleted. Also, the links between  $n_1$  and  $n_2$  and between  $m_1$  and  $m_2$  are added to the graph. If the nodes do not have the condition then the new node should be chosen.

By applying algorithm 3.1, the average clustering coefficient for the new graph is computed, and it will be compared with the previous average clustering coefficient. If the new average clustering coefficient is smaller than the previous computed average clustering coefficient, then the links of the graph are changed back to the previous situation. If the new average clustering coefficient is higher than the previous average clustering coefficient, it means that the average clustering coefficient of the graph is optimized. This algorithm is continued until the average clustering coefficient of the graph become larger than the average clustering coefficient of the generated Erdos-Renyi random graph, or the algorithm is repeated for the specified times.

By the end of this algorithm, if the average clustering coefficient of the graph becomes higher than the average clustering coefficient of the Erdos-Renyi random graph, then the graph with the Lognormal node degree distribution and the average clustering coefficient higher than the average clustering coefficient of the Erdos-Renyi random graph is generated. On the other hand, if the algorithm 4.1 is stopped since the algorithm is repeated a specified number of times,

then it shows that the average clustering coefficient of the generated graph is not larger than the average clustering coefficient of the Erdos-Renyi random graph. So the new graph should be generated. This process is continued until the graph with higher average clustering coefficient than the average clustering coefficient of the Erdos-Renyi random graph is generated.

#### 4.1.3.2 Average Short Path Length

In the next step, by applying the average short path length algorithm, the average short path length of the generated graph produced in the previous section and the Erdos-Renyi random graph are computed. If the calculated average short path length of the graph is smaller than the average short path length of Erdos-Renyi random graph, then the generated graph has the small world characteristic, also it has the same node degree distribution as Facebook network graph. Therefore, the user network graph of Facebook is simulated. If the generated graph does not have the small world characteristic (has a larger average short path length than the Erdos-Renyi random graph), the new graph should be generated again. This process is continued until the graph with the same node degree distribution and small world characteristic is produced.

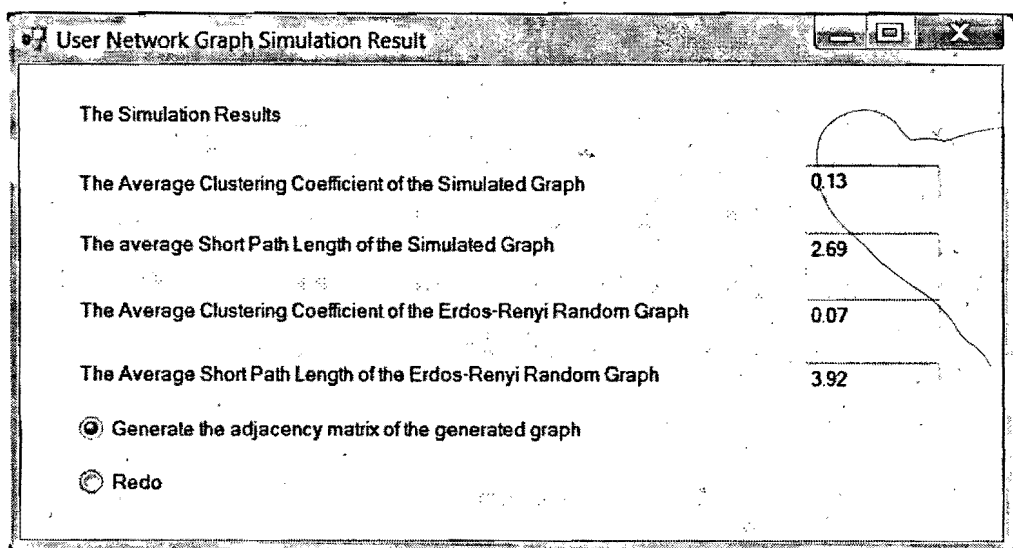


Figure 4.3 User Facebook Network Graph Simulation Result form



Figure 4.3 shows the user interface that can be used by the user to do the visual comparison until getting the appropriate graph. As it is shown, if the simulated graph qualifies for all the conditions the result can be sent to a file in the form of adjacency matrix.

## 4.2 Simulation Graph

Figure 4.4 and figure 4.5 present Facebook user network graph graphically which are simulated by applying the proposed algorithm showed in algorithm 4.1 that is produced by Matlab software.

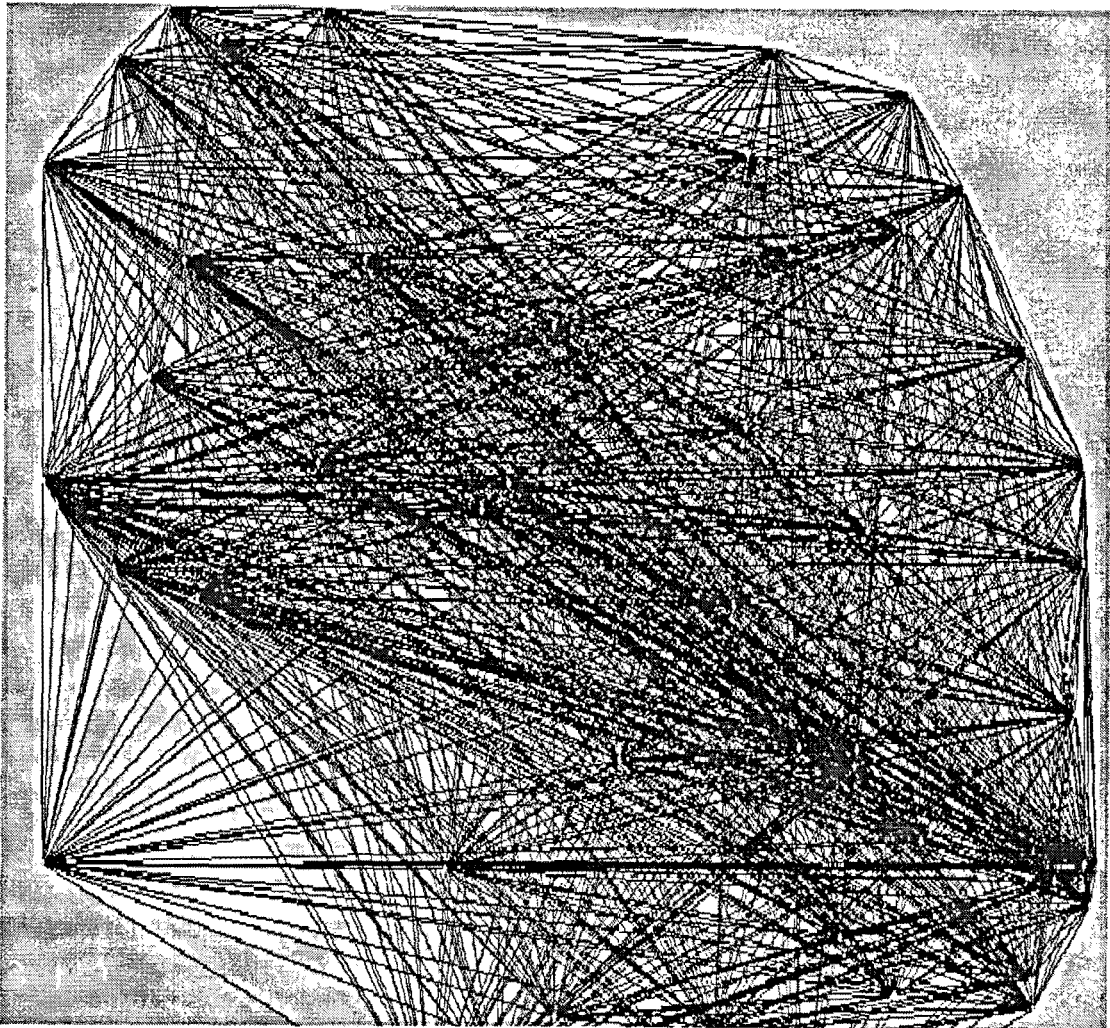


Figure 4.4 Simulated Facebook user network graph for 50 nodes

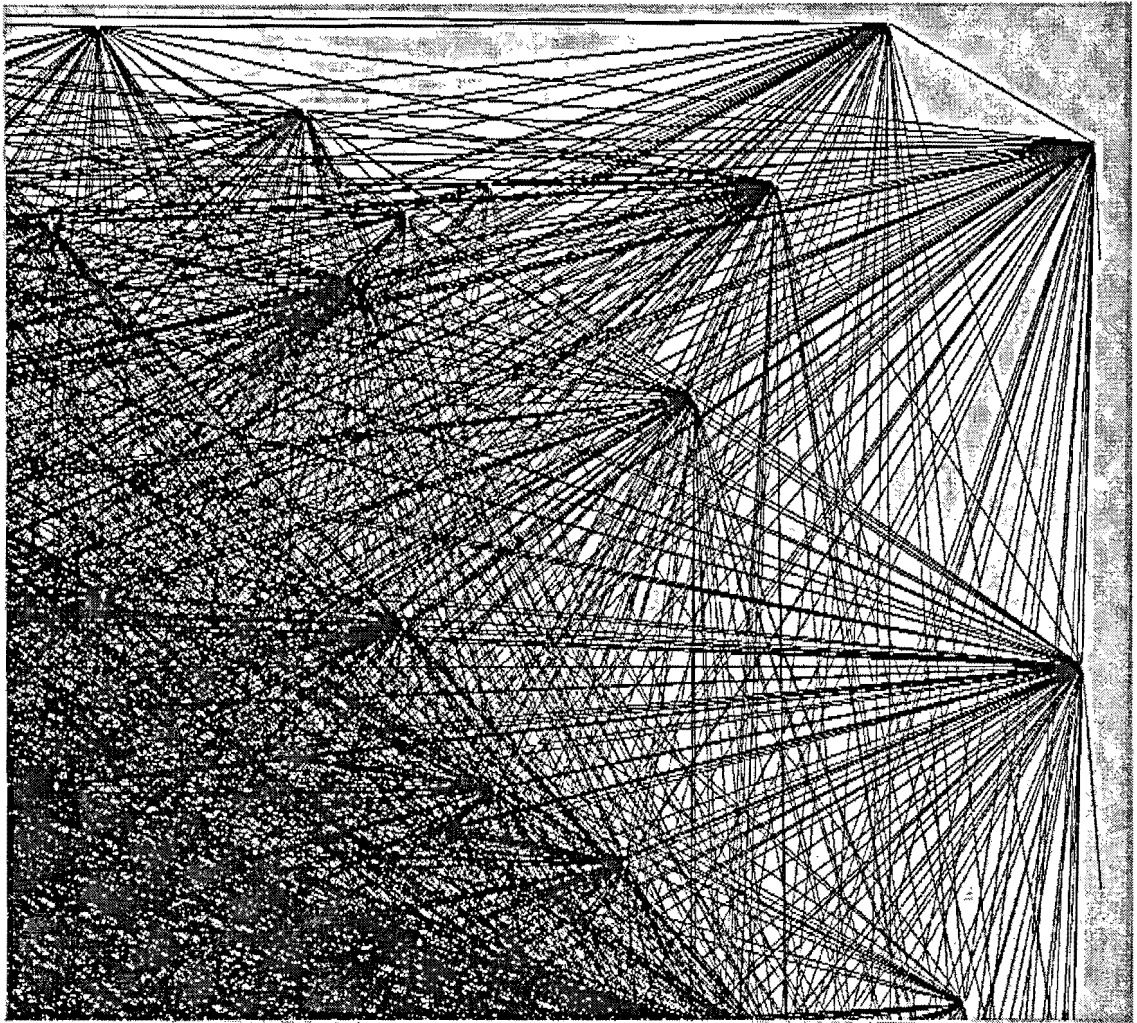


Figure 4.5 Simulated Facebook user network graph for 100 nodes

### 4.3 Generating Web Page Fans Graph

As we mentioned in chapter 3, the Weibull distribution fits the distribution of the increase in the number of different Facebook web pages fans. Therefore, the Weibull distribution is applied to build the software to generate the increase in the number of Facebook web page fans.

Our software is used to generate 200 random data that shows the increase for each day (200 day in total) and the generated data follows the Weibull distribution. Table 4.1 presents the statistics of generated data.

Table 4.1 Summary statistics of the generated data

Characteristic	Generated Data
Mean	1,760.89
Median	1,364
Standard Deviation	1,622.58
Maximum	11742
Minimum	7
Sample Size	200
CV	0.92
Model	Weibull a=1.12 b=1,838.19

To test our software CDF of generated data is compared with Weibull and Gamma distributions. The Kolmogorov-Smirnov (K-S) test is also applied to test the goodness of fit. As table 4.2 and visual comparison shown in figure 4.6 distribution follows the Weibull distribution.

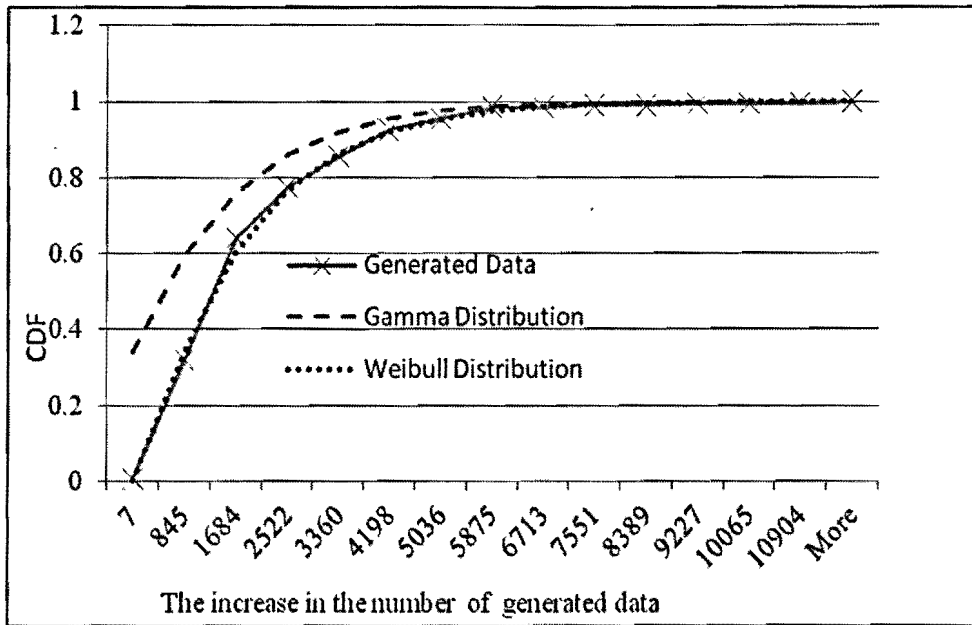


Figure 4.6 The Distribution of the generated data

Table 4.2 Kolmogorov-Smirnov Test

Model	K-S Statistics (Generated Data)
Weibull Distribution	0.024451
Gamma Distribution	0.33619

The above results prove that the data generated by our software follows the Weibull distribution as we wanted that validates our simulator software. We did the same test for the data generated for user networks and also it shows the generated data follows the Lognormal distribution as we expected that also validates our developed software. The user interface forms using for generating web pages fans graph are shown in Figure 4.7 and 4.8.

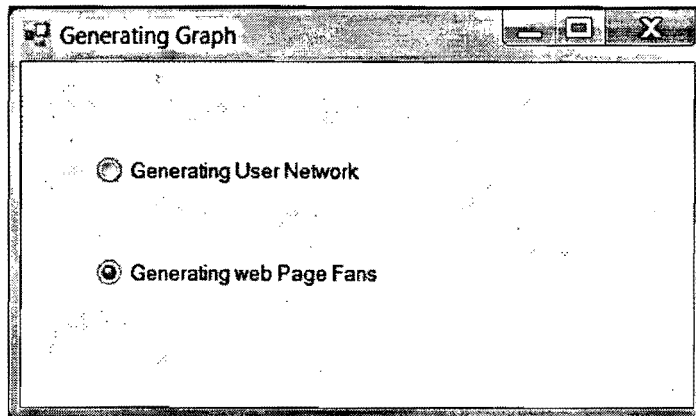


Figure 4.7 Generating Graph

Weibull default value

Scale

Shape

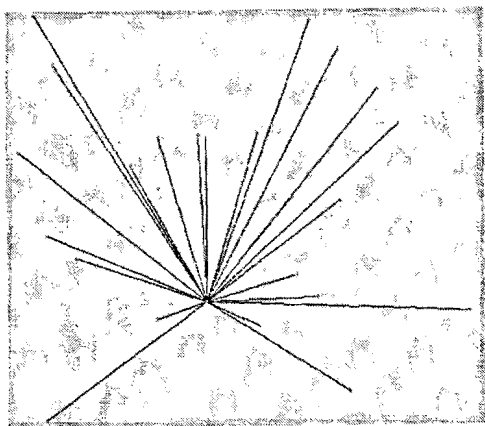
No. Days

Generated Data

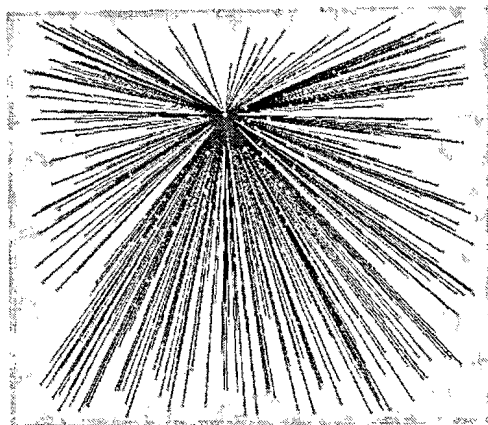
- 120
- 2082
- 2823
- 1391
- 237
- 2953
- 3399
- 63
- 1093
- 2335

Figure 4.8 Generating Web Page Fans form

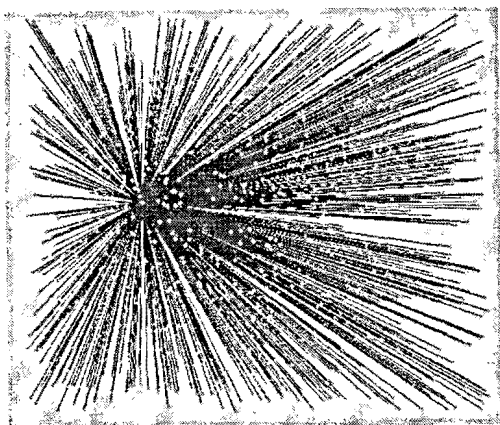
Since web page fan graph is used to model dynamic nature of Facebook, we have run the simulation for 200 days (1 day is corresponding to each data value). Figure 4.9 presents this increase in the number of web page fans graphically in 6 snap shots.



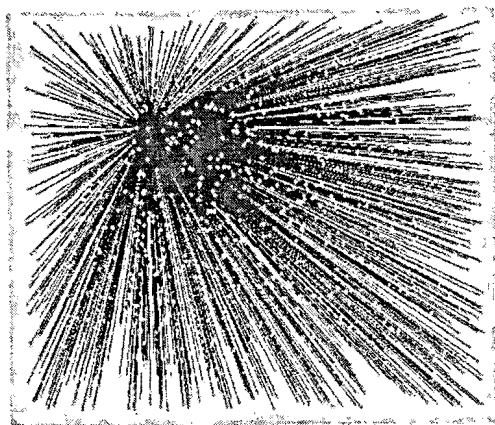
a. The 1<sup>st</sup> day



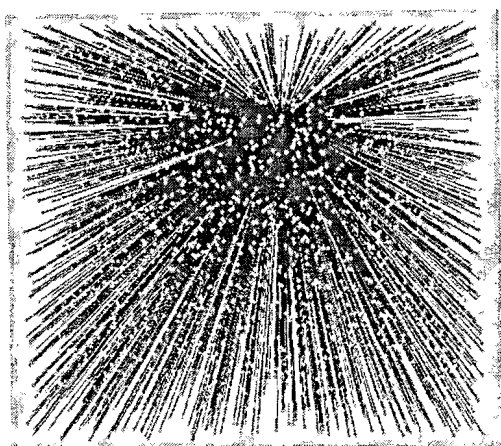
b. The 30<sup>th</sup> day



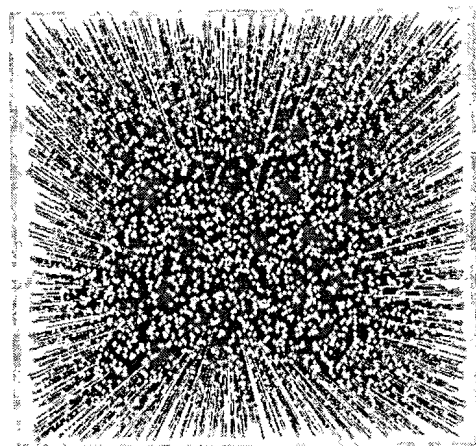
c. The 60<sup>th</sup> day



d. The 90<sup>th</sup> day



e. The 120<sup>th</sup> day



f. The 150<sup>th</sup> day

Figure 4.9 The increase in the number of web page fans during the simulation of 200 days

### 4.3 Conclusions

In this chapter, by applying Facebook user network features that are analyzed in the previous chapter, we simulate the Facebook user network graph. We observed that the user network graph has two conditions: It has the Lognormal node degree distribution and small world characteristic. It has also the higher average clustering coefficient and a smaller average short path length than the Erdos-Renyi random graph.

Also in this chapter, the Weibull distribution is used in the software to generate the increase in the number of Facebook web page fans for 200 days. At the end of this chapter we test our software by comparing the graph of the generated data with the Weibull and Gamma distribution graphs and also by applying the Kolmogorov-Smirnov fitness test. The results validate the correctness of the generated output by the software.

## Chapter 5

### 5. Conclusions and Future Works

This chapter of the thesis presents the conclusions of our research in section 5.1, and section 5.2 presents future works.

#### 5.1 Conclusions

In Facebook a small number of users generate large traffic through the applications installed on their Facebook. An active Facebook user may upload many videos and share them with their friends. This generates large data on storage systems and many requests on servers of Facebook data centers. One of the implications of this research is building the synthetic workload simulator that can be used in the simulations for evaluating methods to reduce traffic of the web 2.0 sites similar to Facebook. To reach this goal, we modeled the user networks and social characteristics of Facebook such as the existence of small world phenomenon. These models can also be used to determine the structure and the speed of data transfer in user networks.

In this modelling we have considered two assumptions as follows: First, since the number of user's friends is changing less frequent compared to the number of web page fans, we assume that Facebook user network has static nature during our measurement interval. Second, since the number of web page fans increases frequently, we measured that to capture the dynamic nature of increase in the number of Facebook member.



The results of our modelling present the characterization of Facebook attributes and its relations to network traffic. Two data sets are collected from Facebook. The first data set is collected from the user profiles during a seven-month period that contains 20,019 Facebook users' name and the list of their friends. The second data set is collected from several web pages belonging to different categories such as marketing companies, TV show, TV channel, political people, and entertainment during a 131-day period. The first data set shows that the node degree distribution of Facebook user network fits the Lognormal distribution. The second data set shows that the distribution of the increase in the numbers of Facebook web pages fans can be modeled by the Weibull distribution.

This thesis also presents the average clustering coefficient and generating Erdos-Renyi random graph algorithms. The first data set is used to estimate the average clustering coefficient of the collected Facebook user network graph. The Erdos-Renyi random graph with the same number of nodes and links is generated as well in this thesis. The comparison between the average clustering coefficient of the Facebook user network graph and the generated Erdos-Renyi random graph presents that Facebook user network graph has higher average clustering coefficient than the Erdos-Renyi random graph.

Also the average short path lengths for the Facebook user network graph and the generate Erdos-Renyi random graph are measured in this thesis. The comparison between these two average short path lengths presents that the Facebook user network graph has smaller average short path length than the Erdos-Renyi random graph. Therefore, the results suggest that Facebook user network has the small world characteristic. It means that any two users in Facebook user network can link to each other by small number of hops.

The models for node degree distribution of Facebook user network and the small world characteristic are used to simulate Facebook user network. Then, the simulator generates a graph that has node degree distribution following the Lognormal distribution. Also the simulated graph has higher clustering coefficient and smaller average short path length than the Erdos-Renyi random graph with the same number of nodes and links.

Also the distribution of the increase in the number of web page fans which is the Weibull distribution is used in the synthetic simulator software which generates the increase in the number of Facebook web page fans (i.e., for dynamic modelling). In this thesis, the software is tested, and the result validates synthetic workload output.

## **5.2 Future Works**

The results of this thesis can be applied for multiple future works. By modelling the user network in Facebook the synthetic workload simulators that were developed can be used in finding the methods to reduce traffic both in Facebook server and network. It can be used in performance measurement tests for the methods addressing network traffic issues.

Second of all, as we know, broadcasting news plays a vital role in nowadays. Marketing companies spend a large amount of money on television or radio commercials. Thus, the resulted simulator software can be used in the experiments regarding to understanding of information flow and data transfer in the Facebook user networks. This thesis presents that the Facebook user network graph has a small world characteristic. In other words, any two users in the Facebook user network graph can link to each other by small number of hops. Thus news can spread faster between Facebook users without spending a large amount of money on commercials and advertisements.

Parametric features of the simulator software can be used to generate different scenarios and structures for the fan's networks and user networks graph in Facebook in order to identify how fast companies can broadcast their promotional events by using Facebook.

All of these experiments can be the future direction for this work that are worthy for doing research.

## References

- [1] P. Nagpurkar, W. Horn, U.Gopalakrishnan, N. Dubey, J. Jann, P. Pattnaik, "Workload Characterization of selected JEE-based Web 2.0 Applications", *In Proceedings of IEEE International Symposium*, Seattle, WA, pp. 109-118, September 2008
- [2] E. Cheng, M. Davis, P. Schmitz, S. Boll, "Web 2.0 and Multimedia: Challenge, Hype, Synergy", *In Proceedings of the 14<sup>th</sup> annual ACM international conference on Multimedia; Panel Session*, Santa Barbara, USA, pp. 752, October 2006
- [3] A. Nazir, S. Raza, C. Chuah "Unveiling Facebook: A Measurement Study of Social network Based Applications", *In Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, Vouliagmeni, Greece, pp. 44-56, October 2008
- [4] Official Web site of Gear6, What is Gear6? <http://www.gear6.com/> last visited October 12, 2009
- [5] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the Evolution of User Interaction in Facebook", *In Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, Barcelona, Spain, August 2009
- [6] M.Gjoka, M. Sirivianos, A. Markopoulou, X. Yang, "Poking Facebook: Characterization of OSN Application", *In Proceedings of Workshop on Online Social networks*, Seattle, WA, USA, August 2008
- [7] X. Kang, H. Zhang, G. Jiang, H. Chen, X. Meng, K. Yoshihira, "Understanding Internet Video Sharing Site Workload: A View from Center Design", *In Proceedings of the 17th international conference on World Wide Web*, Beijing, China, pp. 129-138, April 2008

- [8] Official Web site of Orkut- Online Social Network Web site, <http://www.orkut.com>
- [9] Official Web site of LiveJournal- Express Yourself, Share Your Life, <http://www.livejournal.com/>
- [10] Official Web site of Flickr-Shares your photo, <http://www.flickr.com/>
- [11] P. Gill, M. Arlitt, Z. Li, A. Mahanti, "YouTube Traffic Characterization: A View From the Edge", *In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, San Diego, California, USA, October 2007
- [12] P. Gill, M. Arlitt, Z. Li, A. Mahanti, "Characterizing User Sessions on YouTube", *In Proceedings of SPIE- the International Society for Optical Engineering*, San Jose, CA, USA, January 2008
- [13] X. Cheng, C. Dale, J. Liu, "Statistics and Social Network of YouTube Videos", *In Proceedings of the 16th International Workshop on In Quality of Service (IWQoS)*, Twente, the Netherlands, pp. 229-238, June 2008
- [14] Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee," Measurement and Analysis of Online Social Networks", *In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, San Diego, California, USA, October 2007
- [15] Official Web site of Metacafe- Online Video Entertainment, <http://www.metacafe.com/>
- [16] A. Stuart, K. Ord, S. Arnold, "Classical Inference and the Linear Model. Kendall's Advanced Theory of Statistics", Sixth edition, pp. 25.37-25.43, London: Arnold, 1999

- [17] A. C. Tamhane, D. D. Dunlop, "Statistics and Data Analysis: From Elementary to intermediate", pp. 24-30, pp. 49-52, Prentice-Hall, Inc., NJ, USA, 2000
- [18] S. Milgram, "The Small World Problem", *Psychology Today*, vol.2, no. 1, pp. 60-67, 1999
- [19] R. Albert, H. Jeong, A.-L. Barabasi, "Diameter of the World-Wide Web", *Nature*, vol. 401, pp. 130-131, 1999
- [20] M.E.J. Newman, "The structure of scientific collaboration networks", *Proceeding of the National of the Sciences of the United States of America (PNAS)*, vol. 98, no. 2, pp. 404-409, 2001
- [21] L. A. N. Amaral, A. Scala, M. Barthelemy, H. E. Stanley, "Classes of Small World Networks", *Proceeding of the National Academy of the Sciences of the United States of America (PNAS)*, vol. 97, no. 21, pp. 11149-11152, 2000
- [22] S. Pool, and M. Kochen, "Contacts and influence", *Social Networks*, vol. 1, no. 1, pp. 5-51, 1978/79
- [23] P. W. Holland, S. Leinhardt, "Transitivity in structural models of small groups", *Comparative Group Studies*, vol. 2, no. 2, pp. 107-124, 1998
- [24] D. J. Watts, S. H. Strogatz, "Collective dynamics of 'small-world' networks", *Nature*, vol.393, pp. 440-442, 1998
- [25] J. Bouttier, P. Di Francesco, E. Guitter, "Geodesic distance in planar graphs", *Nuclear Physics*, vol. 663, no. 3, pp. 535-567, July 2003

- [26] E. Horowitz, S. Shahni, "Fundamental of Data Structures", pp. 293-295, W H Freeman & Co, 1983
- [27] X. Cheng, C. Dale, J. Liu, "Characteristics and Potentials of YouTube: A Measurement Study. Peer to Peer Video", in *Peer-to-peer video -the economics, policy and culture of today's new mass medium*, E. M. Noam, & L. M. Pupillo (Eds.), Springer, pp. 205–217, New York, USA, 2008
- [28] M. Maia, J. Almeida, V. Almeida, "Identifying User Behavior in Online Social Networks", *In Proceedings of the 1st Workshop on Social Network Systems*, Scotland, pp. 1-6, 2008
- [29] S. Mitra, M. Agrawal, A. yadav, N. Carlsson, D. Eager, A. Mahanti, "Characterizing Web-based Video Sharing", *In Proceedings of the 18th international conference on World Wide Web*, Madrid, Spain, pp. 1191-1192, April 2009
- [30] F. Pakzad, A. Abhari, "Characterization of User Networks in Facebook" , *In Proceedings of the 13th Communications and Networking Simulation Symposium (CSN10) of SCS/SpringSim '10 in Collaboration with ACM/SIGSIM* ,Orlando, Florida, US., April 2010
- [31] A. M. Law, W. D. Kelton, "Simulation Modelling and Analysis", Third Edition, pp. 347-370, McGraw-Hill, 2000
- [32] A. M. Law, Associates, "ExpertFit© Version 6 User's Guide", February 2004.  
[www.averill-law.com](http://www.averill-law.com)
- [33] F. Pakzad, A. Abhari, "Modelling User Networks in Facebook", *Social Network Analysis and Mining*, New York, Submitted on July 2010

- [34] T.H. Cormen, C.E. Leiserson, R.L. Rivest, "Introduction to Algorithms", pp. 469-477, MIT Press, New York, 1990
- [35] P. Erdos, A. Renyi, "On Random Graphs. I", *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959
- [36] P. Erdos, A. Renyi, "The Evolution of Random Graphs", *Academy Mathematics Institute of Hungary Science*, vol. 5, pp. 17–61, 1960
- [37] A.-L. Barabasi, and R. Albert, "Emergence of scaling in random networks", *Science*, vol. 286, no. 5439, pp. 509-512, 1999
- [38] D. Eppstein, and J. Wang, "A steady State Model for Graph Power laws", *In proceedings of the 2<sup>nd</sup> International Workshop on Web Dynamics*, Honolulu, 2002
- [39] J.M. Kleinberg, "Navigation in a small world", *Nature*, vol. 406, no. 6798, pp. 845, August 2002
- [40] M. E. J. Newman, D. J. Watts, S. H. Strogatz, "Random Graph Models of Social Networks", *Proceeding of the National Academy of Sciences of the United States of America (PNAS)*, vol.99, no.1, pp. 2566-2572, 2002
- [41] W. Guo, S. B. Kraines, "A Random Graph Generator with Finely Tunable Clustering Coefficient for Small-world Social Networks", *In Proceedings of International Conference on Computational Aspects of Social Network*, Fontainebleau, France, pp. 10-17, June 2009