AUTOMATIC DATA EXTRACTION FROM ONLINE SOCIAL NETWORK AND TRACK THE CHANGES IN VULNERABILITY

A Dissertation

Submitted to
The Central Department of Computer Science and Information Technology,
Institute of Science and Technology
Tribhuvan University

In Partial Fulfillment of the Requirements for the degree of
Master of Science in Computer Science and Information Technology

By
Satya Bahadur Maharjan
Roll No:- 14
October 2012

Under the Supervision of
Prof. Dr. Shashidhar Ram Joshi
(IOE,TU)

Co-Supervisor
Mr. Bikash Balami
(TU)

Date:………………..

Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Satya Bahadur Maharjan,** entitled "**Automatic data extraction form online social network and track the changes in vulnerability**" be accepted as fulfilling in part requirements for the degree of Masters of Science. In my best knowledge this is an original work in computer science.

Prof. Dr. Shashidhar Ram Joshi
Department of Computer and Electronic Engineering
Institute of Engineering, Pulchowk, Nepal
(Supervisor)

Date:………………..

Recommendation

I hereby recommend that the dissertation prepared under my co-supervision by **Mr. Satya Bahadur Maharjan,** entitled "**Automatic data extraction form online social network and track the changes in vulnerability**" be accepted as fulfilling in part requirements for the degree of Masters of Science. In my best knowledge this is an original work in computer science.

Mr. Bikash Balami
Lecturer
Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur, Kathmandu
(Co-Supervisor)

Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information Technology

We certify that we have read this dissertation work and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirement of Master of Science in Computer Science and Information Technology.

Evaluation Committee

Assoc. Prof. Dr. Tanka Nath Dhamala
Head, Central Department of Computer Science and Information Technology
Tribhuvan University

Prof. Dr. Shashidhar Ram Joshi
Department of Computer and Electronics Engineering
Institute of Engineering
Pulchowk, Nepal
(Supervisor)

(External Examiner)

(Internal Examiner)

Date: ------------------------

ACKNOWLEDGEMENT

It is my great opportunity to complete this thesis under the supervision of Prof. **Dr. Shashidhar Ram Joshi**, Department of Computer and Electronics Engineering, Institute of Engineering Pulchowk, Nepal. I would like to express my profound gratitude to my supervisor for his generous encouragements and undertaking of the supervision of my entire research work. This form of report is the outcome of his continuous encouragement, helpful suggestions and comments.

With this regard, I wish to extend my sincere appreciation to respected Head of Department of Central Department of Computer Science and Information Technology, **Assoc. Prof. Dr. Tanka Nath Dhamala** for his kind help.

I would like to thank co-supervisor **Lecturer Bikash Balami**, Central Department of Computer Science and Information Technology, Tribhuvan University, Kirtipur, Kathmandu, for their valuable guidance and helpful suggestions.

I am also grateful to all staffs of Central Department of Computer Science and Information Technology, Tribhuvan University and all other staffs of Central Library of Tribhuvan University.

I would like to thank Mr. Umesh Maharjan, the senior programmer and analyst, Hitech Valley inet for his valuable guidance and helps in developing web crawlers.

Lastly, let me offer my profound gratitude to my parents, family members and friends who inspired and helped me to uplift me at every steps of my life. This study is an outcome of their heartily blessing.

ABSTRACT

The popularity of Online Social Network has increased huge amount of personal data to be found on Online Social Network and those data may get vulnerable. This makes people vulnerable to social engineering attacks because their personal data are readily available.

In this research work, an automated data extraction tool is developed and stored them in repository. An online social network graph was generated according to the data stored in repository. Here node represents people's profile. The graph analysis identifies structural features of the node such as indegree, outdegree, degree centrality, closeness centrality, betweenness centrality, and clustering coefficient. These all graph features are calculated, using social network analysis tool i.e. NodeXL, according to likes and comments of the status posted by a user. The node is said to be vulnerable node, if the value of clustering coefficient is towards 1. It depends on the high number of interaction of those friends who have more mutual friends.

The vulnerability of a node may change during the change of time. Timestamp is given for those nodes whose vulnerability changes during the change of time was evaluated again by analyzing the previous status with new updates.

Keyword: Online Social Network, vulnerability, timestamp

LIST OF FIGURES

LIST OF TABLES

CONTENTS

Details **Page No:-**

LIST OF ABBREVIATIONS

SN	Social Network
OSN	Online Social Network
SNA	Social Network Analysis
DEX	Data Exchange
SocNetV	Social Networks Visualizer
Ms – Excel	Microsoft Excel
IM	Instant Messaging
GPS	Global Positioning System
URL	Uniform Resource Locator
GraphML	Graph Markup Language
SNAP	Stanford Network Analysis Platform library
API	Application Programming Interface
SMS	Short Messaging Service

CHAPTER 1
INTRODUCTION

1.1 Social Networking

A Social Networking (SN) is an online network through which one can share ideas, activities, events and interests within their individual networks [1, 4, 12]. These types of network are developed on building strong social relations among people. It is not only used by individual but also by business organizations, celebrities, bands, groups, companies, colleges and schools etc. These are formed to share the knowledge and information.

Most of the schools, colleges, or offices are available with internet. The internet is filled with millions of individuals who are looking to meet other people, to gather and share their information and experiences about their employment, business, marketing, cooking, gardening, programming etc. They may create their groups of common interest in hobbies, religion, politics and alternative lifestyles [4]. Once you become the member of SN, you begin to socialize. This socialization can be performed by managing profile pages as adding friends and removing uncommon friends [17].

With the emergence of this social networking website the world has become closer. Long lost family, schoolmates and friends are able to connect with each other from across the globe.

1.2 History

The starting of Online Social Network (OSN) starts with classmates.com in 1995. The first well known OSN site is SixDegrees.com which was launched in 1997 and closed down in 2000 [24]. It was named after the theory *six degrees of separation* that anyone can be connected to any other person through a chain of acquaintances that has no more than five intermediaries [17, 24]. Through this OSN users could create their profiles, have a list of friends and contribute information to their community.

The table below shows some of the OSN sites' establishment

S.N	**List of OSN**	**Data of establishment**
1	Classmates.com	1995
2	LiveJournal	1999
3	Friendster	2002
4	Hi5	2003
5	Linkedin	2003
6	MySpace	2003
7	Orkut	2004
8	Facebook	2004
9	Flicker	2004
10	Google+	2011

Table 1.1: History of OSN [13]

All the social networking sites are developed not for the specific purpose. Social networking sites such as MySpace, Facebook, Orkut, and LinkedIn are examples of wildly popular networks used to find and organize contacts. Other social networks such as Flickr, YouTube, and Google Video, are used to share multimedia content, and others such as LiveJournal and BlogSpot are used to share blogs.

1.3 Popularity of OSN

Online social networking sites are becoming more popular day by day [8]. The number of running online social networking among other sites are becoming more.

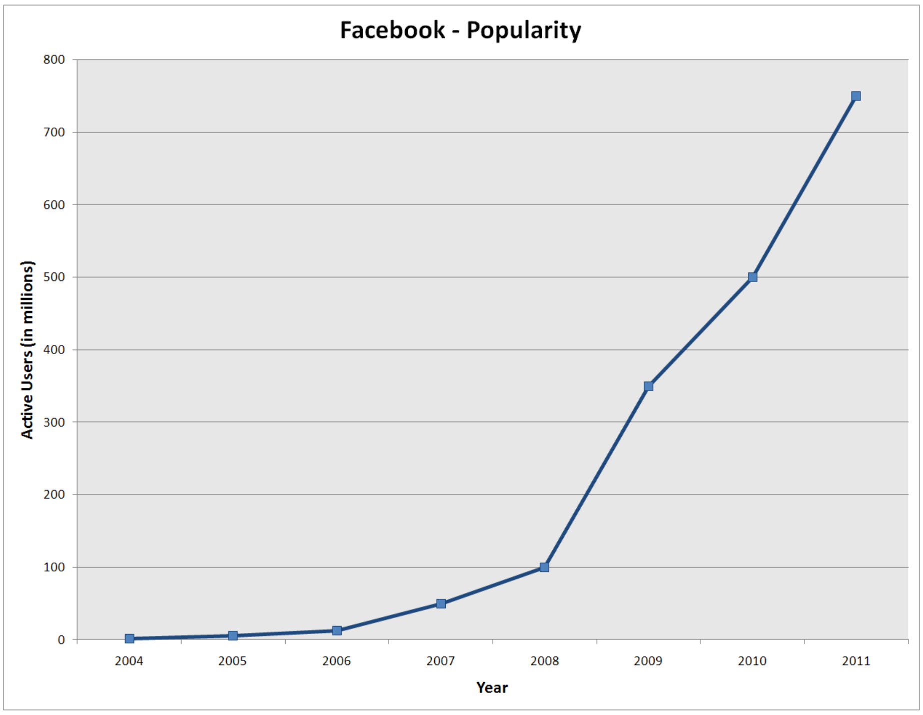


Figure : 1.1 Graph showing the increasing Facebook users on OSN [14]

The running of the OSN is increasing day by day[6]. The traffic of Facebook is increasing day by day and the number of users is about 800 million plus. More than 50% of Facebook active users log on in any given day. The average user has 130 friends. They interact with different pages, groups, events and community pages [8].

Facebook is not only accessed in computer. More than 350 million active users currently access it through their mobile devices. More than 250 million photos are uploaded per day [8].

The popularity of online social networking sites has increased the amount of personal data which is distributed on the net [3]. These profiles are semi-structured and the profile data or structure may be changed in an unpredictable way [3, 7].

Facebook and other networking tools is increasingly the object of scholarly research [3, 10]. Scholars in many fields have begun to investigate the impact of social networking sites, investigating how such sites may play into issues of identity, privacy, social capital, youth culture, and education. If the structure of Facebook is changed, then tool must be modified.

1.4 How Information Stored in OSN May Leak Our Data

Anyone can create a profile in which they include personal data and information. OSN provide different tools to interact with others. The main features of a OSN and their tools are: Communication (allow sharing knowledge), Community (help finding and integrating communities), and Cooperation (provide tools to develop activities together) [15]. So, the information stored on OSN is readily available to the user and with these information user can used for social engineering attacks, which allowed malicious website to access the real name and also they could post bogus messages on their wall [3].

These events have highlighted the issue of privacy in OSN's and how personal details can get without control into the wrong hands. Displaying personal details can make users more vulnerable to social engineering attacks like identity theft and re-identification by linking. These would enable people to extract personal details from the OSN profiles and use external sources to find out more about that person's identity [2].

1.5 Impact of OSN on society

Those information that are posted on OSN can be viewed by any users through online. That's why the data stored on OSN may be vulnerable. Some of the researchers found that the data stored is published publicly [4].

1.6 Friend

Any users can join a network, publish their own content, and create links to other users in the network called "friends" [18]. This basic user-to-user link structure facilitates online interaction by providing a mechanism for organizing both real-world and virtual contacts, for finding other users with similar interests, and for locating content and knowledge that has been contributed or endorsed by "friends".

The extreme popularity and rapid growth [8] of these online social networks represents a unique opportunity to study, understand, and leverage their properties. Not only can an in-depth understanding of online social network structure and growth aid in designing and evaluating current systems, it can lead to better designs of future online social network based systems and to a deeper understanding of the impact of online social networks on the Internet. Online social networks also contain many useful properties that can be leveraged to enhance information systems, such as enhancements to controlling information propagation, new directions for information search and retrieval, and new ways of reasoning about trust.

1.7 Social Network Analysis software

There are much more software which can analyze the data extracted from different social networks [6]. Those software are Social Network Analysis (SNA) software which has the features of describing features of a network either through numerical or visual representation. For eg: Automap (text mining tool), DEX (Data Exchange) (graph database for query processing and network analysis), graph-tool (python module for efficient analysis and visualization of graphs), NodeXL (Network Overview Discovery Exploration for Excel), Social Networks Visualizer (SocNetV)etc.

1.8 NodeXL

The open source software tool, NodeXL was designed especially to facilitate learning the concepts and methods of SNA with visualization. It is a template for Microsoft Excel (Ms-Excel) 2007/2010 Add-in and C#/.Net library for network analysis and visualization. It supports directed graphs as well as non-directed graph. Network relationship are represented as an edge list which contains all pairs of vertices that are connected in the network. Users are allowed to display the range of network graph representation and map properties like shape, color, size, transparency, and location. It allows for easy manipulation and filtering of underlying data in spreadsheet format [11].

This thesis is focused on 8 distinct chapters as: chapter 2 covers vulnerability on OSN from different factors. In chapter 3, the problem is formulated with the solution, the objective of the study and literature is reviewed. Chapter 4 covers research strategy and experimental setup. Chapter 5 covers data preparation and chapter 6 covers analysis and discussion. In chapter 7, Conclusions and finally chapter 8 covers future studies and limitations.

CHAPTER 2

VULNERABILITY OF ONLINE SOCIAL NETWORK

2.1 Vulnerability

Vulnerability means a weakness which allows an attacker to reduce a system's information assurance [3, 12, 15]. That means disclosure of data from OSN. Because of its popularity and the number of hits per day is becoming more and more from its research data stored on OSN may be vulnerable [2,8].

2.2 Vulnerable Node

The vulnerable node in a social network graph is the node that contains the attributes to breach privacy and provide grounds for a social engineering attack. Also the node has a highly connected neighbourhood in which the neighbours display the attributes readily [3].

For example, in November 2007, when users used Beacon (the advertising system that monitored Facebook users) for shopping, Facebook shared the data to their friends. Hackers can use the personal data available on OSN profiles and other public resources to gain access to online systems like email and banking systems. Those data can be used to reset the online systems.

This section introduces vulnerabilities commonly exploited by attacks seeking users' private information.

2.3 The Vulnerabilities Associated to the Platforms of OSN

The vulnerability of data stored in OSN may be through the platform. It may be from:

- **Difficulty to Completely Remove All Users' Information When Deleting an Account**

When any one user tries to leave a social network, there will be the license agreement to those contents that are uploaded. So, if anyone wants to remove those uploaded materials like videos or photos, remove it one by one manually [12,15];

Those photographs or videos posted by others, which do not belong to the user, cannot be deleted by himself. For that, the user can report the contents as inappropriate and for owner or the OSN to remove the material [15].

- **Weak Authentication Method**

The most important vulnerabilities on the web environment is weak authentication methods. The combination of user name and password is misused by the user who seeks easy to remember login details [15].

- **Non Validation of User's Data During Registration Process**

While creating new users in OSN, they do not use a validation process. They just check a valid e-mail address, the preferred validation requirement for the user [15]. Facebook contain sufficient method for the validation of user like sending the confirmation code of the user, through phone calls or using short messaging service (SMS) etc.

2.4 Vulnerability Associated with the Data

The data posted on OSN may get vulnerable. There are some applications like SNAG, ProfileLinker which can read/write access to several OSN accounts [12].

- **Disclosure of Navigation Data**

The information of user like operating system, browser, IP address, mobile systems etc can be available on OSN. Those data may be vulnerable. [12,15].

- **Information Disclosed By the User Status**

Instant Messaging (IM) programs and many other OSN applications provide information about users where about [12, 15]. The phishing of IM may get data. For example, if the user status is off-line in that period of time it usually is on-line the attacker knows that something unusual is going on. This also leads, to providing attackers an easy way to exploit previously found vulnerabilities when the user is away from the computer.

2.5 The Vulnerabilities Associated with the Posted Photographs and Multimedia

The photographs and multimedia posted on OSN may be vulnerable from:

- **Tagging By Others**

One of the most useful features on social networks is tagging [12, 15, 16]. Tagging is way of including the name of the user to that content and are visible to those friends who are the friends of tagged one. But, this feature provides an easy way to find all the photograph which may creates embarrassing or inappropriate ones.

- **Implicit Information with Multimedia Contest**

Most of the OSN allow the uploading of multimedia material. Users make frequent use of this feature, but they are unaware that the content upload contains additional meta-data. This meta data provides details such as the camera with which the photograph was taken, where it was taken (through Global Positioning System- GPS coordinates) or when it was taken, nevertheless, if the user removes the meta-data, several algorithms allow discovering, based on recognizable elements in the picture, the place where the photo was taken [12, 15].

Furthermore, facial recognition systems allow identifying a person on a large amount of photographs. These algorithms combined with other technologies, allow finding singular persons on OSN with an acceptable accuracy [12].

CHAPTER 3
PROBLEM FORMULATION AND LITERATURE REVIEW

3.1 Problem Statement

In OSN, any user can join and create their account with their personal information and profile photo. Through which he can organize his account sending request to his known/unknown friends as well as accepting the request send by friends. He can join different groups according to his interest. In this way, the size of friends may increase.

According to the size, the posted data or information may be spread in their network [21]. They can interact with each other by posting messages and other means.

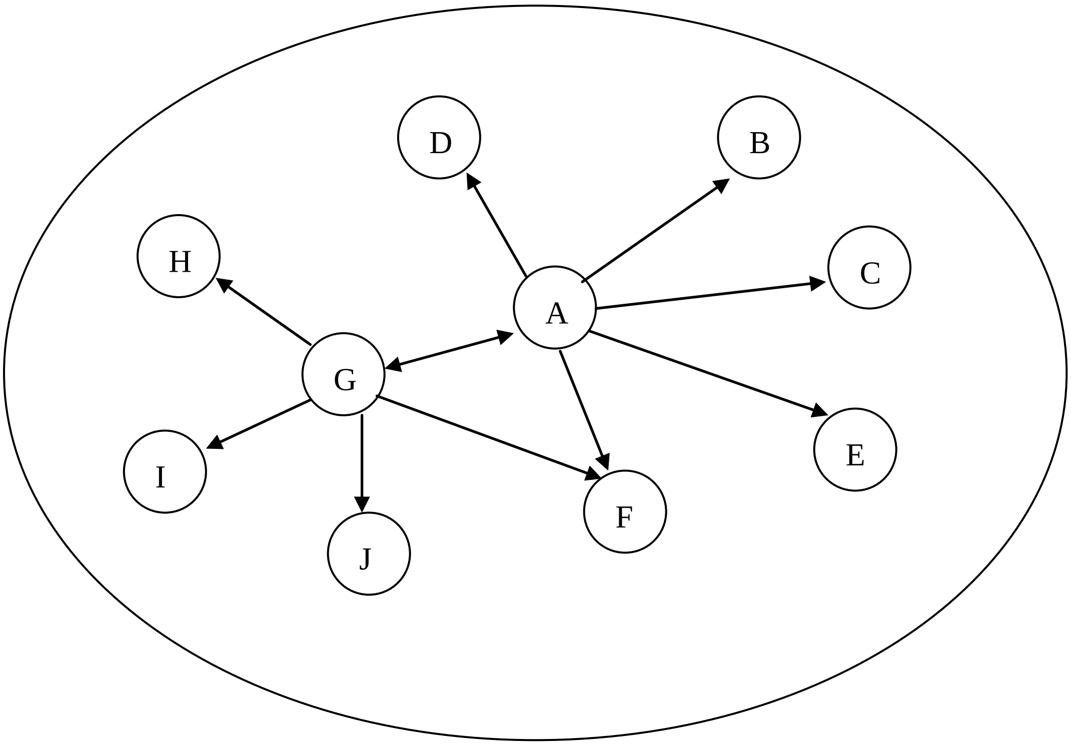


Figure 3.1 Sample graph of friends

Illustrating with the graph that A has 6 friends B, C, D, E, F and G. G also has 5 friends H, I, J, F and A. Here directed graph is created with the flow of messages as A post certain status in his wall that is visible to his 6 friends. If A's friend G comments or likes the posted status of A then that status and A's information is visible to G's friends. Therefore the information posted on A's profile may be vulnerable.

3.2 Graph

A graph is an abstract representation of a set of objects where some pairs of the objects are connected by links. The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges. Typically, a graph is depicted in diagrammatic form as a set of dots for the vertices, joined by lines or curves for the edges.

A graph is an ordered pair G = (V, E) comprising a set V of vertices, people or nodes together with a set E of edges or social relationship between those friends.

The vulnerability of a user is tested using a network analysis software "NodeXL" with visualization "Fruchterman-Reingold" graph is generated [6,11].

When analyzing OSN graphs, there are many characteristics that can be examined. Vulnerability involves the state of the network and the state of the node. For the state of the network, we have to see the average clustering coefficient and the average path length values [3].

3.3 Graph Characteristics

The main characteristics to explore include degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and clustering coefficient.

- **Degree Centrality**

Degree centrality defines the number of links incident upon a node. The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network. In case of a directed network, we define two separate measures of degree centrality: indegree and outdegree.

Indegree means the number of the edges directed towards a node and is denoted by deg (+n). Outdegree means the number of edges directed away from the node and denoted by deg(-n).

Node degree is defined as:

$$\deg(n) = \deg(+n) + \deg(-n) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

- **Closeness Centrality**

The closeness centrality of a vertex measures how easily other vertices can be reached from it. It is defined as the number of vertices minus one divided by the sum of the lengths of all geodesics from/to the given vertex.

$$\mathrm{C_c(i)} = \frac{\mathrm{N-1}}{\sum_{\mathrm{j}} \mathrm{d(i,j)}} \qquad \text{……………………..(2)}$$

As closeness increases, an individual’s access to information, power, prestige etc increases.

- **Betweenness Centrality**

The betweenness centrality of a vertex is the number of geodesics going through it. If there are more than one geodesic between two vertices, the value of these geodesics are weighted by one over the number of geodesics.

$$\mathrm{C_B(n)} = \sum_{\mathrm{s \neq v \neq t \in V}} \frac{\sigma_{\mathrm{st}}(\mathrm{n})}{\sigma_{\mathrm{st}}} \qquad \text{……………………..(3)}$$

Where s and t are nodes in the network different from n, σ_{st} denotes the number of shortest paths from node s to node t and $\sigma_{\mathrm{st}}(\mathrm{n})$ is the number of shortest paths from s to t that n lies on.

The betweenness value for each node n is normalized by dividing by (N-1)(N-2) for directed graph.

- **Clustering Coefficient**

Clustering coefficient of a node reflects how well connected the node's neighbourhood is. We are creating the directed graph. So, the equation of clustering coefficient of node n using equation will be [3, 17, 22]:

$$C_n = \frac{e_n}{(k_n(k_n-1))} \qquad \text{…………………………..(4)}$$

Where $\mathrm{e_n}$ is the number of edges that exist between the neighbours of node n and $\mathrm{k_n}$ is the number of neighbours of node n.

If the value of the clustering coefficient is heading towards 1, then most of the neighbours of a node are connected to each other. On the other hand, if the coefficient value is near 0 then the neighbours are not connected to each other at all [3].

- **Average Clustering Coefficient**

Average clustering coefficient for all the nodes in the network tell us how well connected or not the nodes are to each other which is shown by given equation [3]:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \qquad \text{..............................(5)}$$

Where n is the number of nodes and C_i is the clustering coefficient for each node.

- **Average Path Length**

Average path length of a network will reflect how good or bad the information flow is. It is worked out by averaging all the shortest (geodesic) distances between node pairs of the network. Let graph G have a set of nodes V. The notation for the shortest distance between two nodes is d(v1, v2) where v1 and v2 ∈ V

The equation for the average path length of graph G would be [3]:

$$P_G = \frac{1}{n*(n-1)}\sum_{a,b} d(v_a - v_b) \qquad \text{.................................... .(6)}$$

Where n is the number of nodes in the network and $d(v_a\text{-}v_b)$ is the shortest distance between two nodes.

3.4 Objective

The objectives of this research work are:

- To access the profile data automatically from Online Social Networking site.
- To check the graph characteristics of a user profile.
- To check the vulnerability of a user profile.
- To track the changes in vulnerability.

3.5 Literature review

ALIM S. et al. [2] has shown that how far social network extraction has come since the days when extracting attributes involved a lot of interaction with the profile owner e.g. questionnaires and interviews. Automatic extraction of attributes is the way forward and it can happen with semi-structured web pages. The main challenge when carrying out the experiment to implement the approach was that social networks like MySpace have more than one profile structure template and the user can customise the template [2].

Awareness of the privacy setting on OSN is researched by MOHTASEBI A. and BORAZJANI P. N. on two different universities of Malaysia. They consider different dimensions of privacy and find out to what extent members of those sites are vulnerable to different social attacks [18]. In this paper, there is an attempt to consider different dimensions of privacy and find out to what extent members of those sites are vulnerable to different social attacks. Based on a sample set recruited for this survey, they want to find a relationship between unique characteristics of respondents and their level of vulnerability and awareness about privacy enhancement enablers available in OSN [18].

In this research [18], they used questionnaires and threat model. In questionnaires, basic information with the privacy of the Facebook was collected.

The threat model's objective was to find how many of our targets are vulnerable to our attacks and how easy it is to access private information of a connection. We created a fake Facebook account and tried to access the main information of a targeted contact. We considered an attack a successful one if the targeted contact accepted our fake Facebook user friendship request and it has at least one of the following criteria:

- Access one item of contact information (i.e. Email and IM, Address, Phone Number),
- Access home or work address,
- Access employer name,
- Access personal albums (not profile pictures)

These attacks' intention is to answer the following questions:

- Whether the targets add unknowns persons or not,
- Whether there is a significant effect on the target's decision if she/he has some mutual friends with the fake account.

From threat model analysis approximately 70% of them were accepted without any questions and 30% were rejected or are still pending [18].

Online Social Network grows in size, the practicality of crawling the entire graph decreases. YE S., LANG J. and WU F. investigate the influences of seeds, sample size, node selection algorithms, and the graph being crawled. Author believe that the implications they concluded here shed light on future OSN studies, which will increasingly rely on crawled sub graphs [23].

In this [23] research paper four factors are evaluated:

- **Choice of seed**: Seed is the starting point of a crawling. It is important to select proper seeds.
- **Node selection algorithm**: Node selection algorithms decide which node to crawl next. Some of the node selection algorithm used in this search were breath-first search, greedy algorithm, lottery and hypothetical algorithm.
- **Protected users**: There are concerns that are more and more users adopt the access controls to protect their social data.
- **Different OSNs**: Different OSNs have their unique properties even they provide similar services. In this paper, four OSN Flicker, LiveJournal, Orkut and Youtube were used for research.

CATANESE S. et al. [6] research showed that the data is anonymous and organized as an undirected graph. They described a set of tools that they developed to analyze specific properties of such social-network graphs, i.e., among others, degree distribution, centrality measures, scaling laws and distribution of friendship. In this search work two sampling algorithms: Breadth First Search (BFS) and uniform sampling are used. The methodology implemented to collect data is:

- Preparation for the execution of the agent.
- Starting or resuming the process of data extraction.
- The crawler execution extracts friend lists, cyclically.
- Raw data are collected until the extraction process concludes or it is stopped.
- Data cleaning and de-duplication of information.

Eventually, data structured in Graph Markup Language (GraphML) format.

The Stanford Network Analysis Platform library (SNAP) [6] is used to analyse the dataset collected.

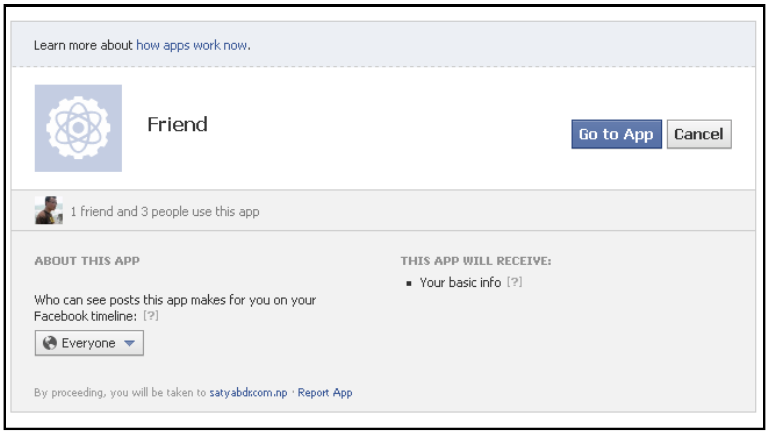
ALIM S. et al. [3] researched on the data extraction from online social network. The aim of this study was to extract all friends from MySpace profiles in order to generate a friendship graph. The graph was analysed to investigate and apply node vulnerability metrics. In this paper only the extraction of top friends but did not investigate the graph or node vulnerability. The graph was generated from the friendship links.

CHAPTER 4

RESEARCH STARTEGY

This architecture is based on the algorithm proposed by Alim S., Abdul-Rahman R., Neagu D. and Ridley M. [3]. The automatic data extraction tool is generated. If the structure of web page is changed the automatic data extraction tool must be modified [7, 9].

First of all specify the URL address to access the data. All the users have unique address through which the users are accessed. The specified URL address is the seeding on which we are going to check the vulnerability of a user [18]. There, we must create a Facebook application (app) through which we can access the friend list. The application also contains app ID and app secret to activate. The application looks like



Learn more about how apps work now.

Friend

1 friend and 3 people use this app

ABOUT THIS APP

Who can see posts this app makes for you on your Facebook timeline: [?]

Everyone

THIS APP WILL RECEIVE:

- Your basic info [?]

By proceeding, you will be taken to satyabdr.com.np · Report App

Figure 4.1: The Facebook application created named Friend.

When we post certain status on the wall of Facebook, there is also unique id for that post.

The proposed system has following major components:

i) **Generating session key**:Session key is the key used to connect in between users for authentication and authorization. The session key must be generated to access the data according to the Facebook application.

ii) **Facebook apps**: Facebook apps must be accepted by all the friends to access their friend list.

iii) **Extracted list of friends**: Extract the list of friends and store in the repository according to the likes and comments of status.

iv) **Generate the OSN graph**: From the repository directed graph is created to study the vulnerability of a node.

v) **Timestamp** the change in repository: Vulnerability of user may change during the change of time. The number of friends may be increased or decreased due to change of time and the vulnerability of the user may be changed and timestamp the change in repository

4.1Working Principle I

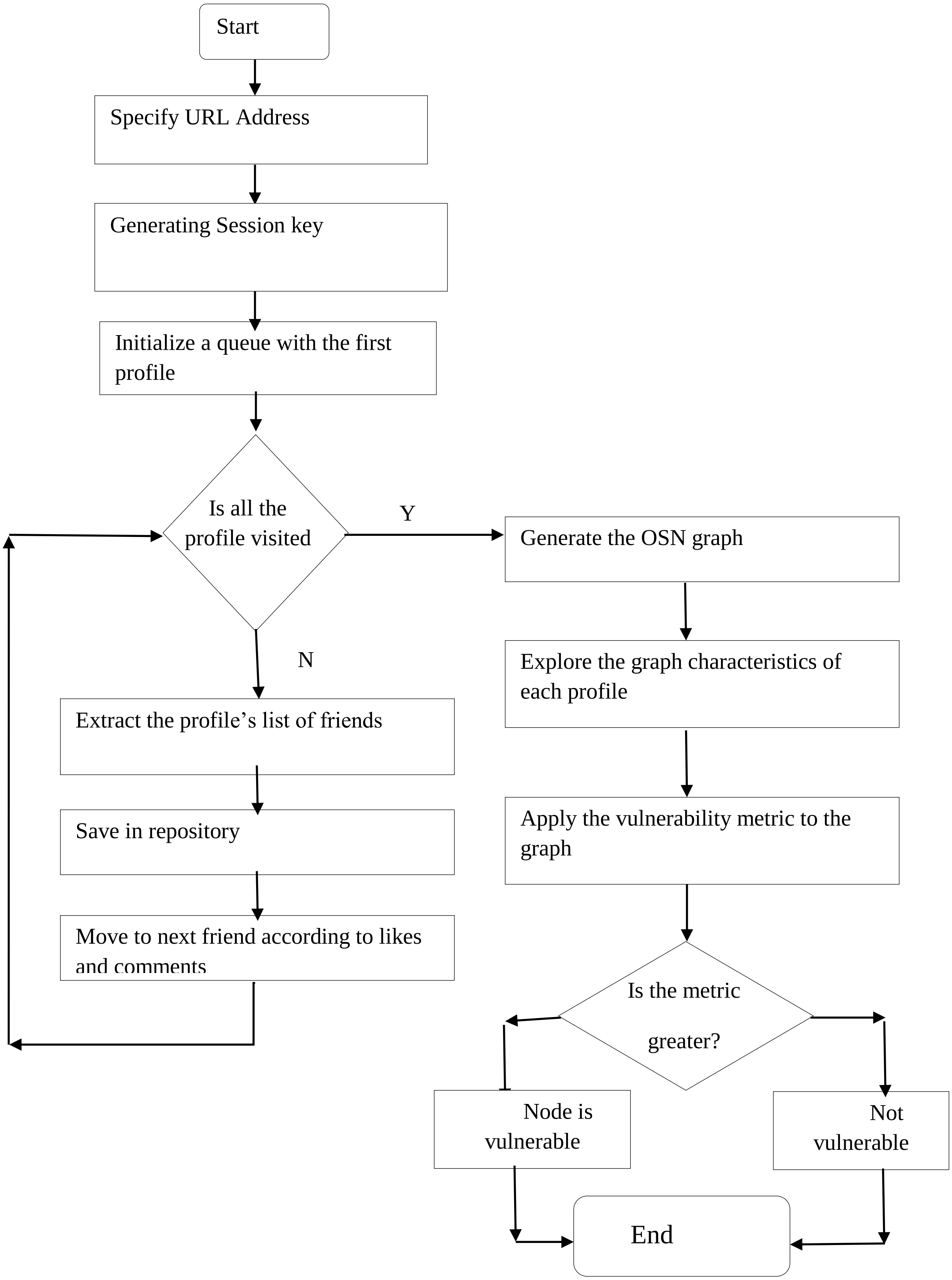

Start
Specify URL Address
Generating Session key
Initialize a queue with the first profile
Is all the profile visited
Y
Generate the OSN graph
N
Extract the profile's list of friends
Save in repository
Move to next friend according to likes and comments
Explore the graph characteristics of each profile
Apply the vulnerability metric to the graph
Is the metric greater?
Node is vulnerable
Not vulnerable
End

Figure: 4.2. Architecture overview – Automatic data extraction and creating graph for vulnerability analysis [3].

4.2 Working Principle II

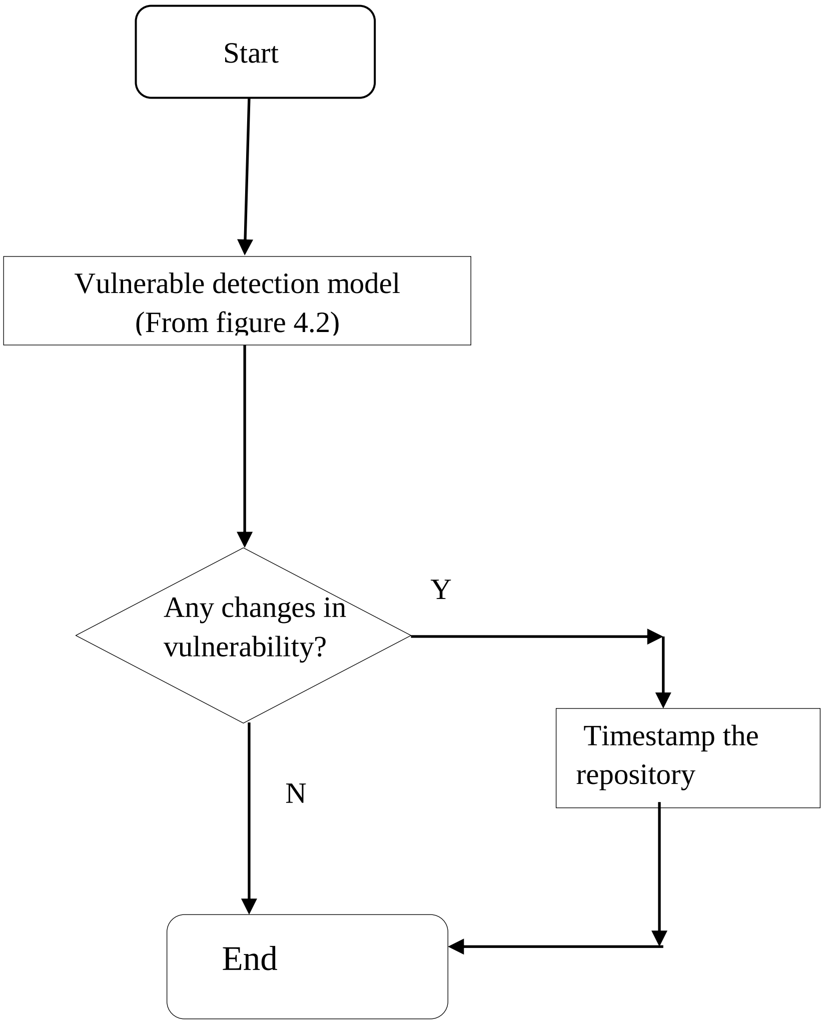


Figure 4.3 Architecture overview- Automatic data extraction, creating graph and track the changes in vulnerability

4.3 Implementation Details

First of all specify the URL address to test the vulnerability analysis of a user. Take a status id of user using Facebook Application Programming Interface (API) with application id and application secret of application friend [fig 4.1]. Generate the session key according to that app. Start to extract the friend list of the specified user and store in repository. Check whether the status has comments or likes. If the status has comments or likes then extract all the friend list of the commentator or liked one and store in the repository. After extracting the entire friend list, generate the OSN graph and apply the vulnerability metric i.e. clustering coefficient. If the size of clustering coefficient exceeds then set the node is vulnerable otherwise set the node as not vulnerable.

For timestamp, after certain time take all those previous users, to check the changes in vulnerability. The value of clustering coefficient may get change i.e. the vulnerability may be increase or decrease and timestamp is given which tracks the changes in vulnerability.

CHAPTER 5

DATA PREPARATION

To test the model, the following data are taken through online via Facebook API. When anyone becomes the member of Facebook s/he has given a unique user id and according to this id, status posted and commented are given the id which becomes unique. For testing data I have collected some user with their id, name, no of friends and id of status posted.

S. N.	**UserID**	**Name**	**No of friends**	**Status ID**
1	664459642	Satya B. Maharjan	313	664459642_10151139674899643
2	100002621473768	Satya Maharjan	25	100002621473768_283039578460089
3	693172287	Shova Shrestha	64	693172287_10151198878917288
4	100003664832586	Kamal Magar	5	100003664832586_151146845017476
5	100003965457610	Ram Shrestha	6	100003965457610_141418002667047
6	597936296	Roshan Silwal	495	597936296_ 10151064733286297
7	703092463	Pravakar Ghimire	360	703092463_10151201009407464
8	572922644	Suresh Thapa	392	572922644_ 396759193723821
9	100000211283456	Bimal Kc	197	100000211283456_509701482380223
10	100001064897384	Arjun Giri	167	100001064897384_443670605678432

Table: 5.1 Data preparation

CHAPTER 6

ANALYSIS AND DISCUSSION

Here the Facebook application is prepared using php and automatic data extraction tool is developed using .Net programming language to access the Facebook data. Besides these NodeXL, the OSN analysis tool, is used to analyze a node.

6. 1 Testing Closeness Centrality

S.N.	**Name**	**Indegree**	**Out-degree**	**Closeness Centrality**	**Max. Closeness Centrality**
1	Satya B. Maharjan	4	313	1.826	3.533
2	Satya Maharjan	4	25	1.922	2.919
3	Shova Shrestha	3	65	1.961	3.042
4	Kamal Magar	2	5	1.760	1.960
5	Ram Shrestha	3	6	1.455	2.727
6	Roshan Silwal	3	495	1.646	3.293
7	Pravakar Ghimire	4	360	1.728	3.149
8	Suresh Thapa	7	392	1.891	3.333
9	Bimal Kc	2	197	1.698	2.744
10	Arjun Giri	4	167	1.848	3.227

Table 6.1: Calculating closeness centrality and its parameter

Central nodes are likely more influential. They have greater access to information and can communicate their opinions to others more efficiently. They are more likely to use the communication channels than other nodes.

6.2 Testing Betweenness Centrality

S.N.	Name	In-degree	Out-degree	Betweenness centrality	Maximum Betweenness Centrality in that network		
					Name	Betweenness centrality	Max Out-degree
1	Satya B. Maharjan	4	313	0.670	Raj Maharjan	1.000	933
2	Satya Maharjan	4	25	0.133	Satya B. Maharjan	1.000	304
3	Shova Shrestha	3	65	0.047	Sumendra Maharjan	1.000	1405
4	Kamal Magar	2	5	0.006	Satya Maharjan	1.000	25
5	Ram Shrestha	3	6	1.000	Ram Shrestha	1.000	6
6	Roshan Silwal	3	495	0.857	Ramesh Naupane	1.000	579
7	Pravakar Ghimire	4	360	0.657	Roshan Silwal	1.000	495
8	Suresh Thapa	7	392	0.320	Sumendra Maharjan	1.000	1405
9	Bimal KC	2	197	0.689	Satya B. Maharjan	1.000	313
10	Arjun Giri	4	167	0.689	Arjun Kumar	1.000	460

Table 6.2: Calculating betweenness centrality with its parameters

Betweenness centrality is based on communication flow. A person (node) who lies on communication flow can be calculated by this graph characteristic.

6.3 Testing Clustering Coefficient

S.N.	Name/User	No of Vertices	No. of Edges	In-degree	Out-degree	Clustering Coefficient
1	Satya B. Maharjan	1762	1855	4	313	0.0008
2	Satya Maharjan	321	344	4	25	0.0233
3	Shova Shrestha	1676	1800	3	65	0.0130
4	Kamal Magar	26	36	2	5	0.3500
5	Ram Shrestha	12	20	3	6	0.1330
6	Roshan Silwal	1402	1544	3	495	0.0005
7	Pravakar Ghimire	1325	1575	4	360	0.0020
8	Suresh Thapa	1165	1311	3	392	0.0006
9	Bimal Kc	654	677	2	197	0.0004
10	Arjun Giri	654	677	3	167	0.0002

Table 6.3 Calculating clustering coefficient and its parameters

Clustering coefficient defines how much the node is vulnerable in that graph. The node whose clustering coefficient is towards 1, is more vulnerable. The visualization of above data is presented in directed graph as:

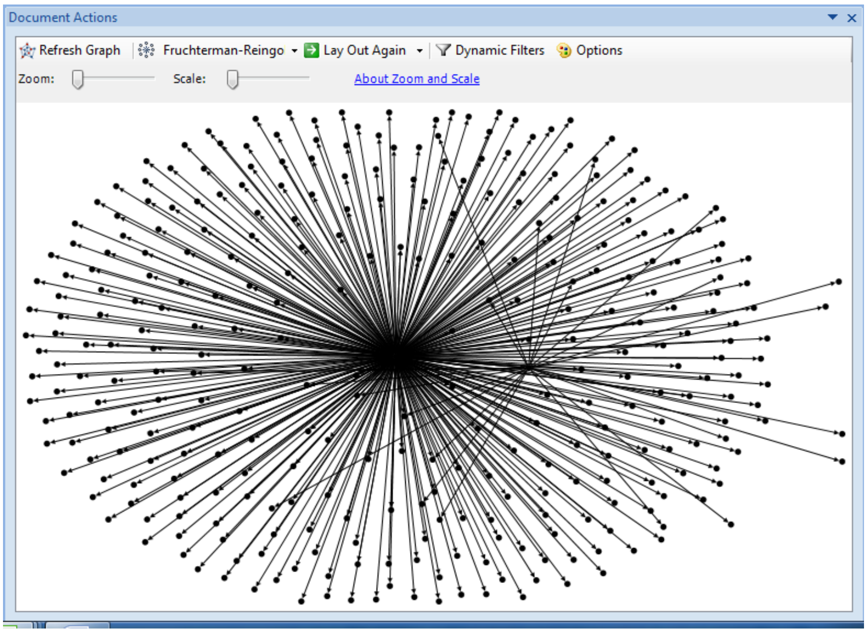


Figure [6.1.a]: Directed graph of 664459642

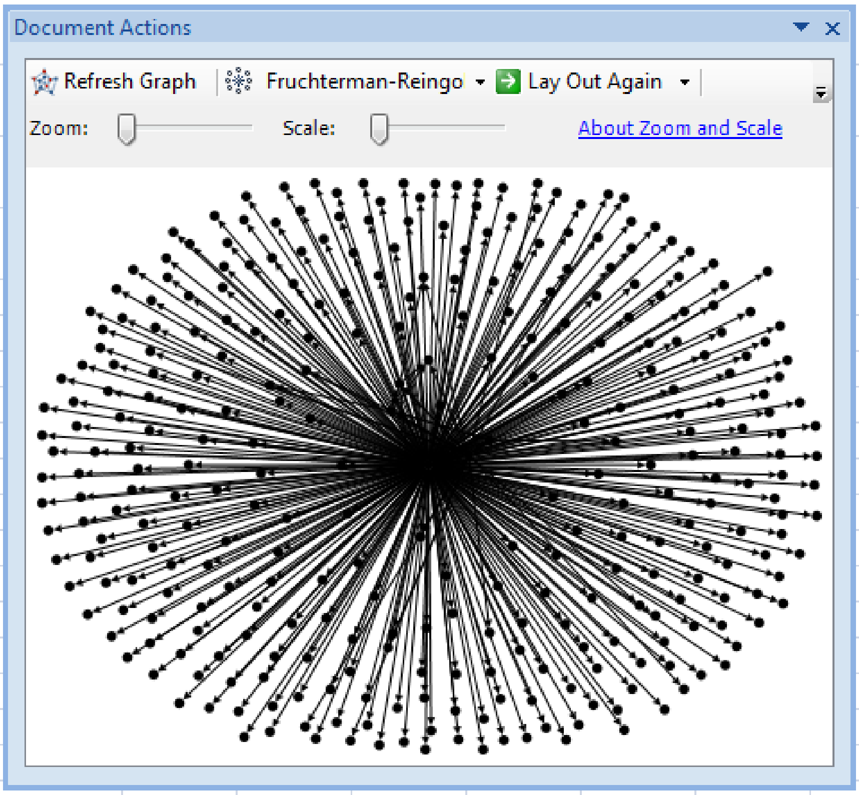


Figure [6.1.b]: Directed Graph of 100002621473768

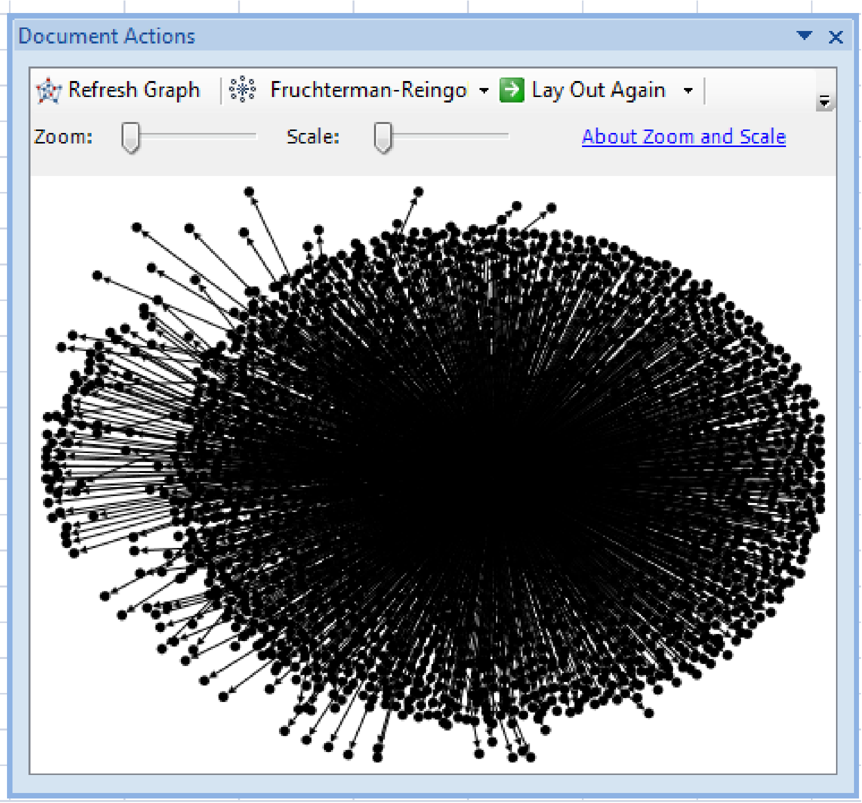


Figure [6.1.c]: Directed graph of 693172287

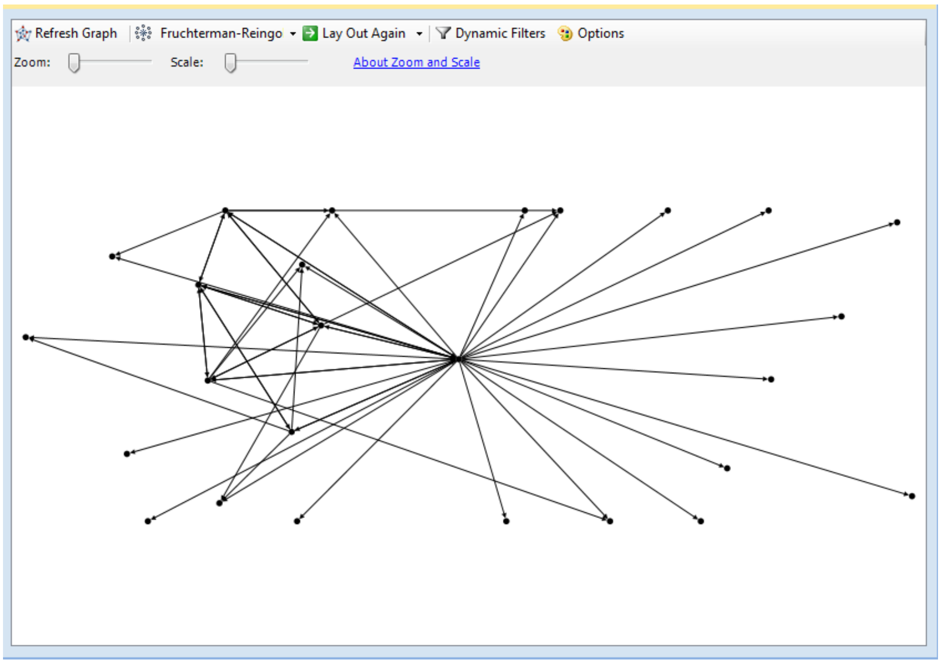


Figure [6.1.d]: Directed graph of 100003664832586

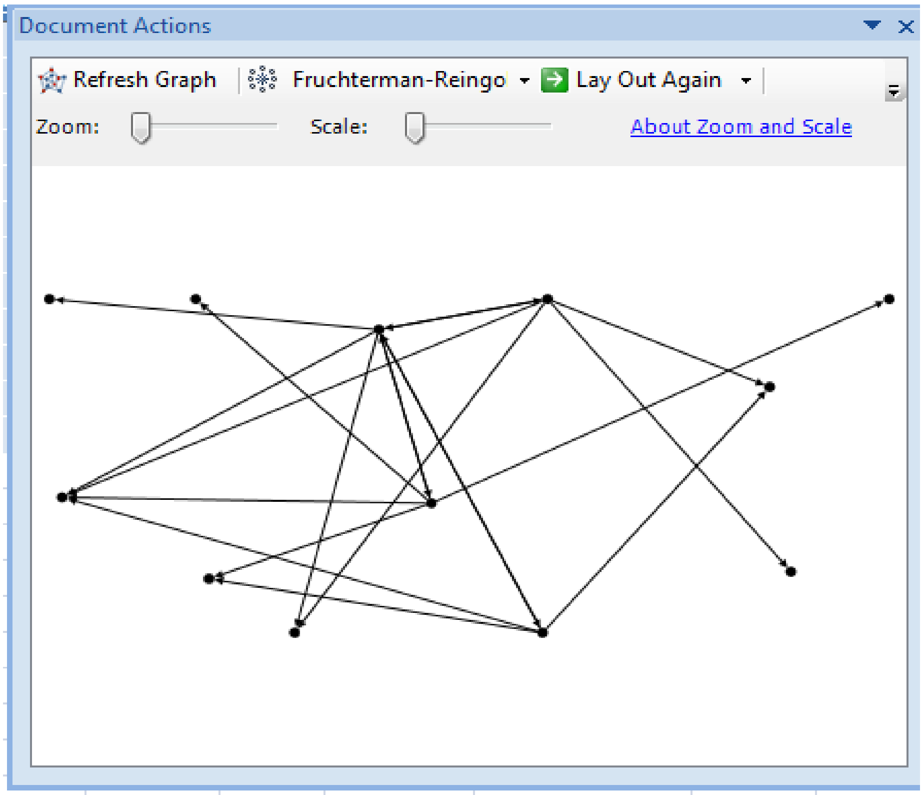


Figure [6.1.e]: Directed graph of 100003965457610

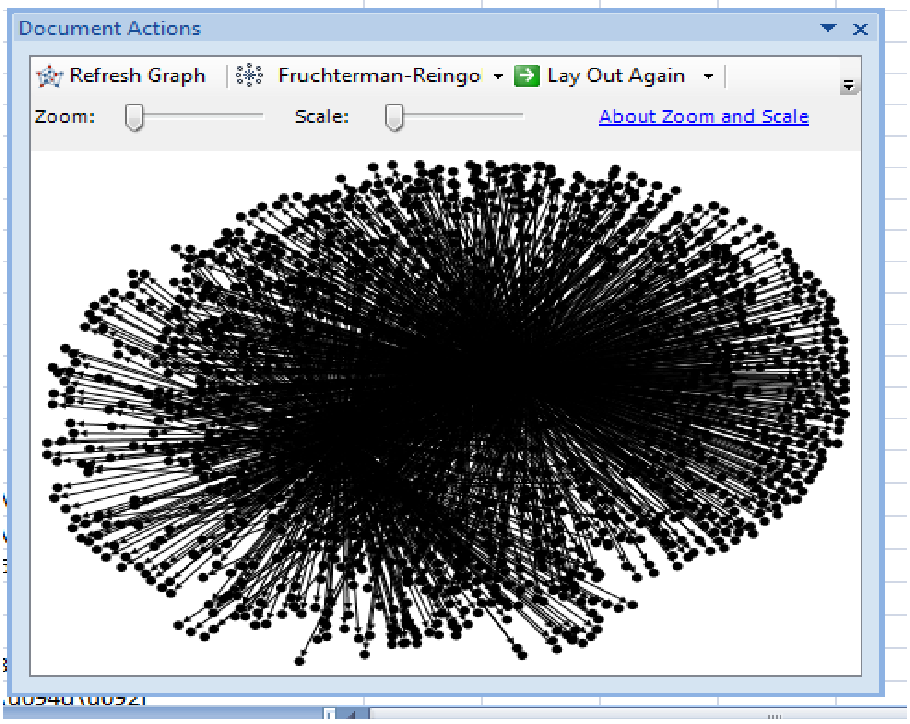


Fig [6.1.f]: Directed graph of 597936296

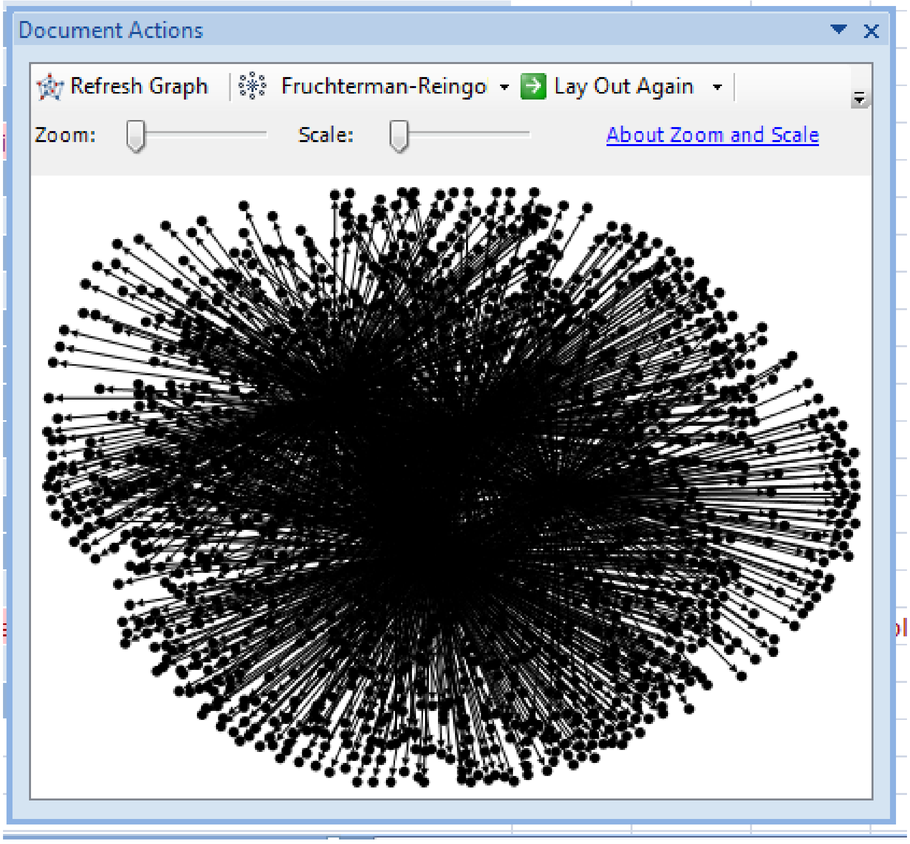


Figure [6.1.g]: Directed graph of 703092463

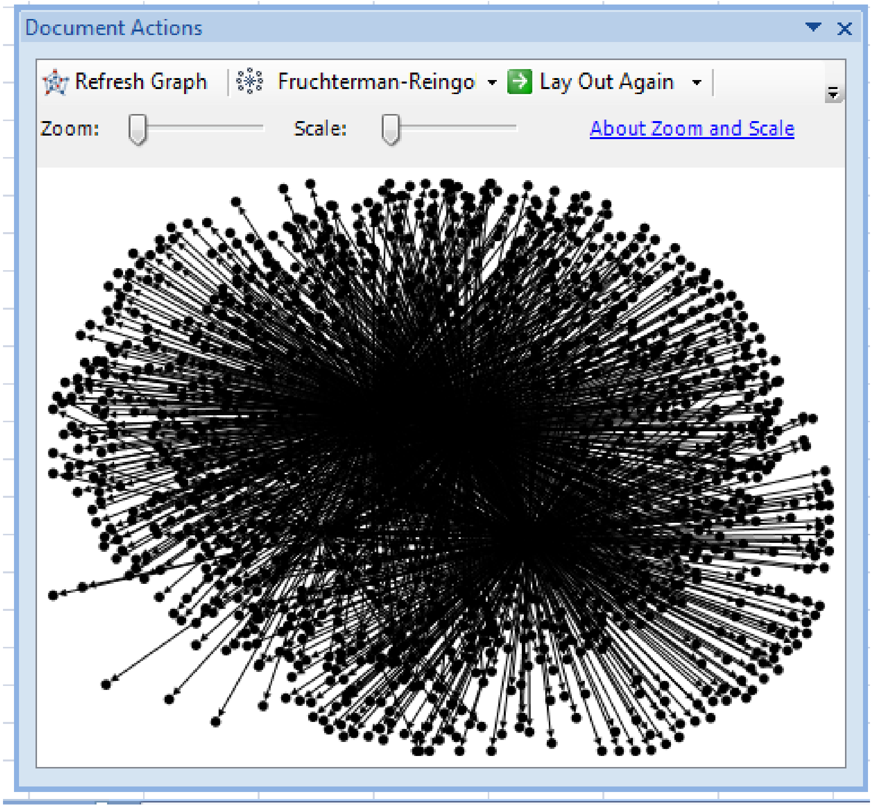


Figure [6.1.h]: Directed graph of 572922644

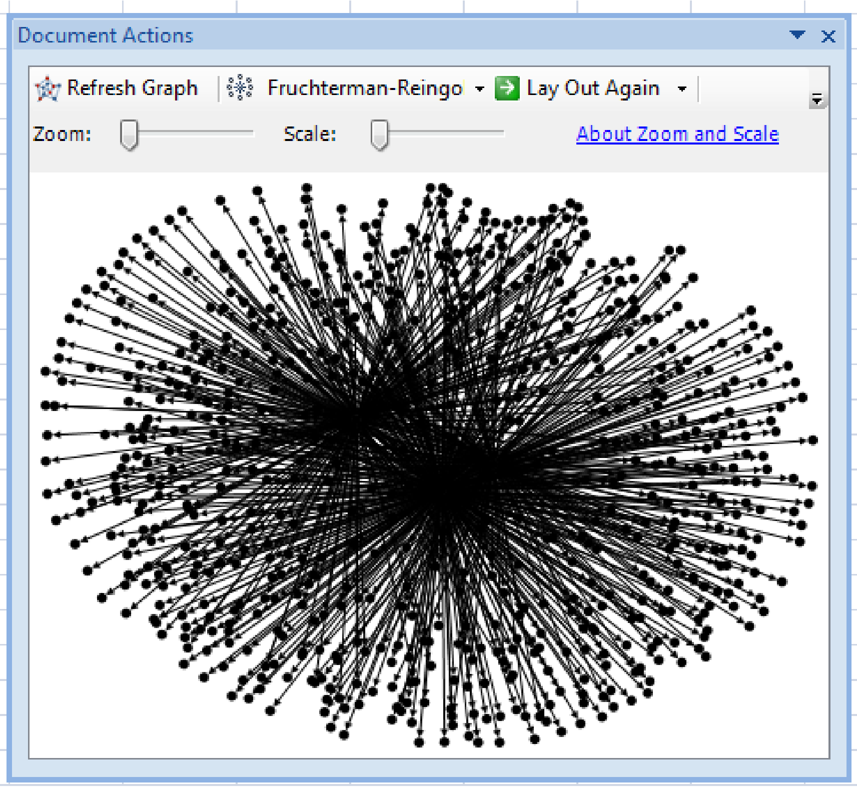


Figure [6.1.i]:Directed graph of 100000211283456

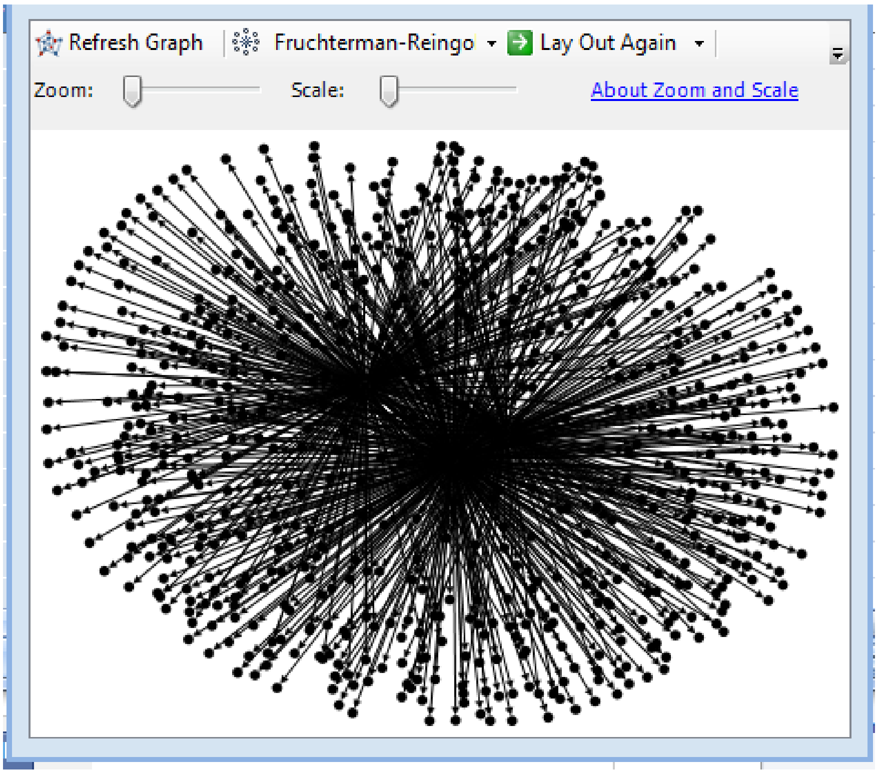


Figure [6.1.j]: Directed graph of 100001064897384

Figure 6.1: Directed graph which shows the flow of messages (from fig a-j)

The data of Facebook may be changeable and due to this property the vulnerability also may get changed. The timestamp of 1 week is taken to track the changes in vulnerability.

6.4 Timestamp After One Week

S.N.	Name/User	No of Vertices	No. of Edges	In-degree	Out-degree	Clustering Coefficient	Changes in vulnerability
1	Satya B. Maharjan	1903	2041	6	313	0.0010	Increased
2	Satya Maharjan	366	422	8	25	0.0400	Increased
3	Shova |Shrestha	2876	3361	7	65	0.0361	Increased
4	Kamal Magar	26	52	4	5	0.6000	Increased
5	Ram Shrestha	26	55	6	6	0.4000	Increased
6	Roshan Silwal	1673	1912	5	495	0.0008	Increased
7	Pravakar Ghimire	1562	1855	5	360	0.0019	Decreased
8	Suresh Thapa	3609	4221	7	392	0.0023	Increased
9	Bimal Kc	654	677	2	197	0.0004	Constant
10	Arjun Giri	1103	1137	4	167	0.0006	Increased

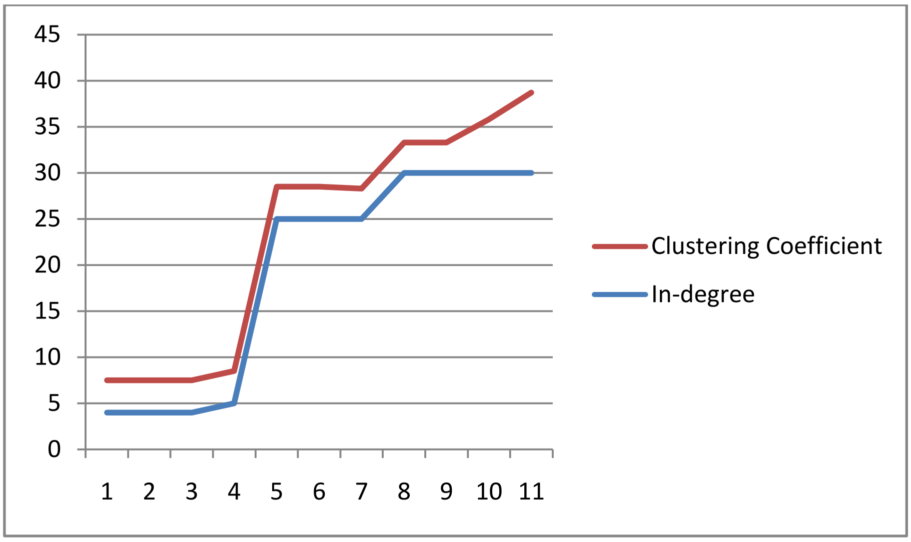
Table 6.4: Calculating clustering coefficient and its parameters after 1 week.

6.5 Empirical Data:

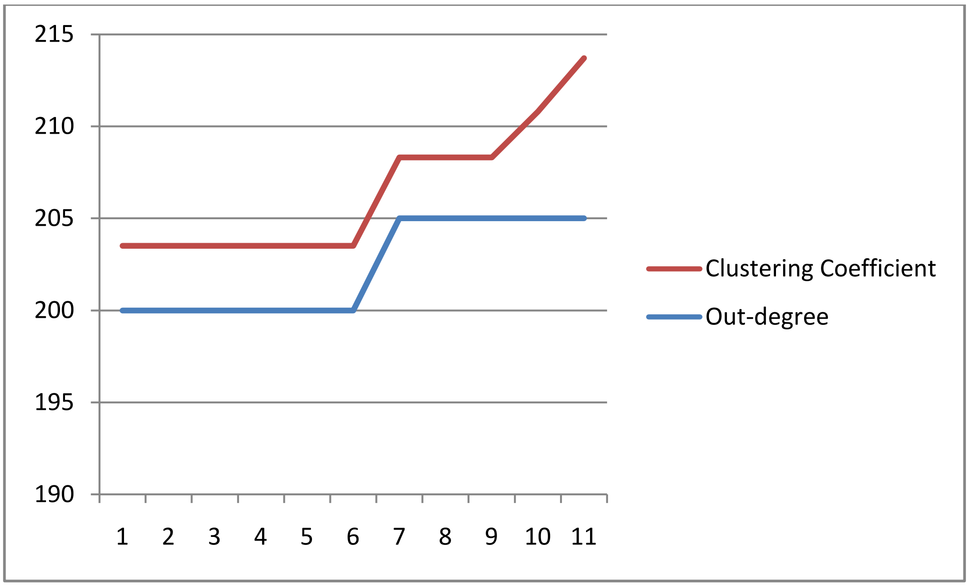
s.n.	Name/User	In-degree	Out-degree	No of vertices	No of Edges	Mutual friends with bimal	Total mutual friends	Clustering Coefficient
1	Bimal Kc	4	200	657	682	12	21	0.00035
2	Bimal Kc	4	200	667	692	12	21	0.00035
3	Bimal Kc	4	200	667	702	12	31	0.00035
4	Bimal Kc	5	200	667	723	12	51	0.00035
5	Bimal Kc	25	200	667	743	12	71	0.00035
6	Bimal Kc	25	200	697	773	12	71	0.00035
7	Bimal Kc	25	205	702	778	12	71	0.00033
8	Bimal Kc	30	205	702	783	12	76	0.00033
9	Bimal Kc	30	205	752	883	12	176	0.00033
10	Bimal Kc	30	205	753	892	22	186	0.00058

Table 6.5 : Emperical data testing of a Facebook user (100000211283456)

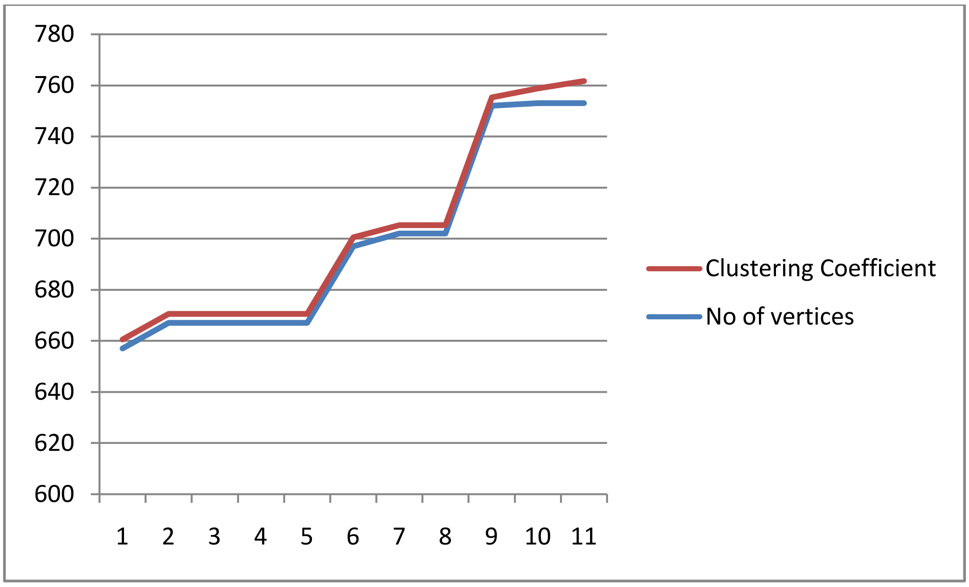
The visualization of table 6.5 is given below:



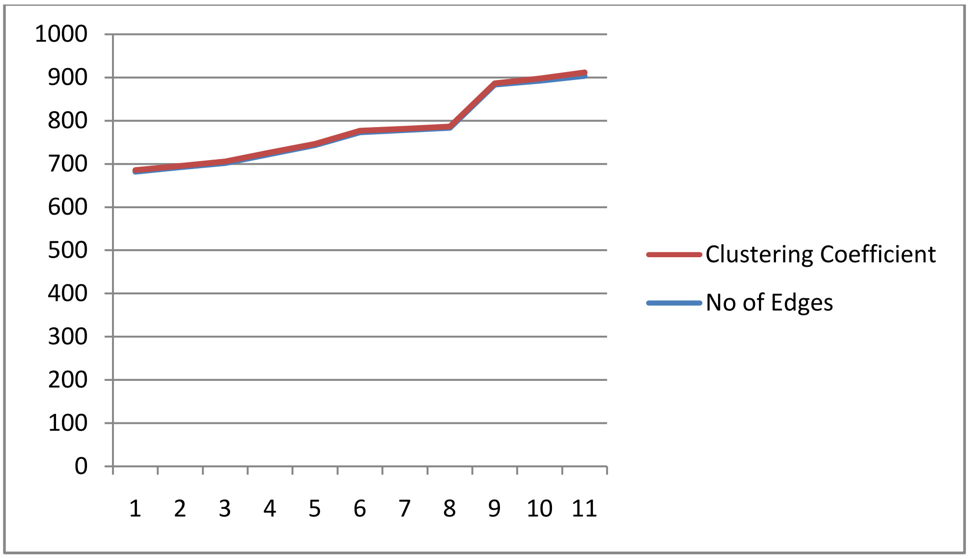
Figure[6.2.a]: Relation between clustering coefficient and in-degree



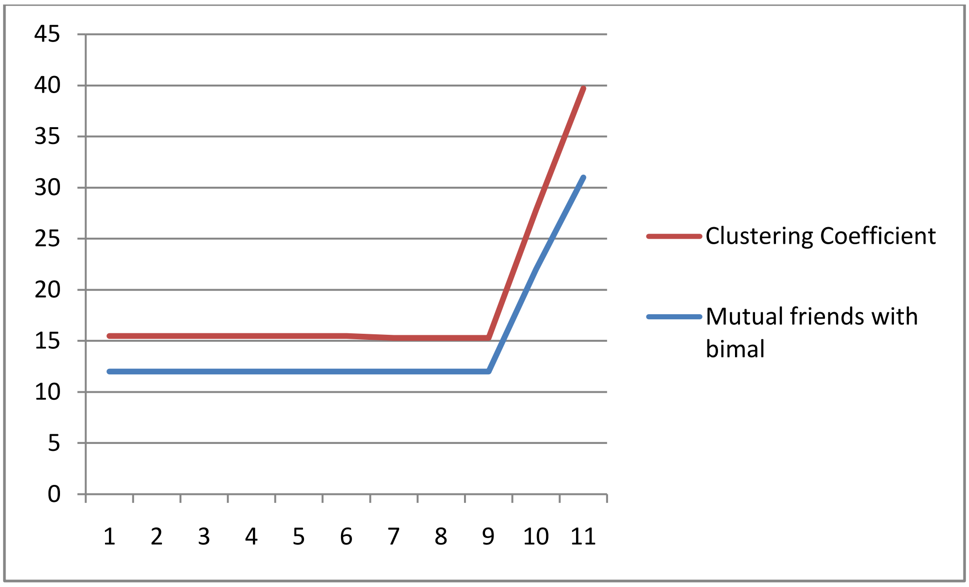
Figure[6.2.b]: Relation between clustering coefficient and out-degree



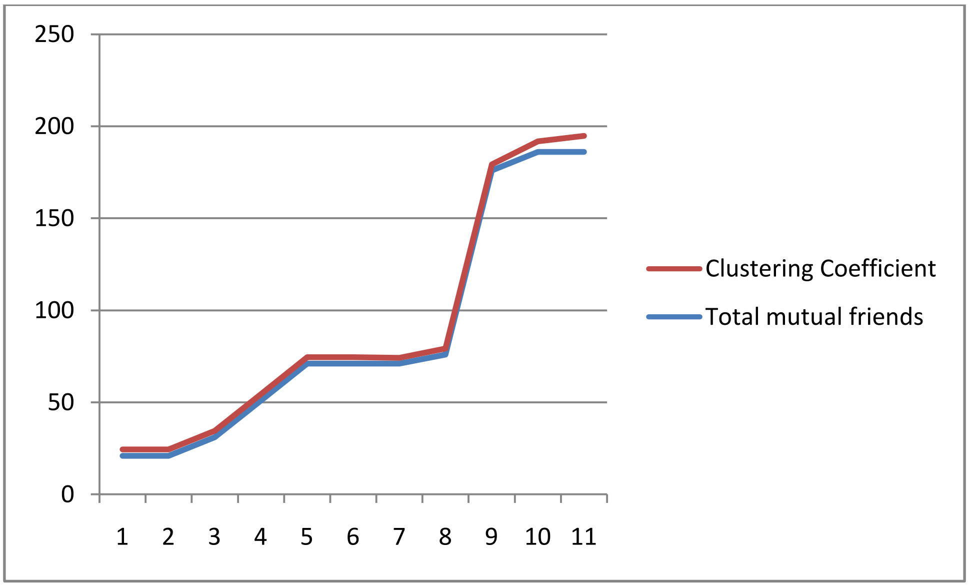
Figure[6.2.c]: Relation between clustering coefficient and no of vertices



Figure[6.2.d]: Relation between clustering coefficient and no of edges



Figure[6.2.e]: Relation between clustering coefficient and mutual friends with bimal



Figure[6.2.f]: Relation between clustering coefficient and total no of mutual friends

Figure 6.2: Relation showing clustering with different parameters from figure[6.2.a] – figure [6.2.f]

From the above empirical data experiment it shows that when the mutual friends increases the vulnerability will be more and also according to the response i.e. in-degree.

CHAPTER 7

CONCLUSION

An OSN site i.e. Facebook is taken for the study and different data having different number of friends are taken for research. Through which we analyzed using network analysis tool NodeXL. Those data are tested using different graph characteristics like indegree, outdegree, degree centrality, closeness centrality, betweenness centrality, clustering coefficient. This showed the flow of information within the network.

The vulnerability of a node is calculated using clustering coefficient. The vulnerability of a node is directly proportional to the number of interactions of user having more mutual friend.

$$vulnerability \propto number\ of\ interations\ of\ user\ having\ more\ mutual\ friend$$

The timestamp is given for one week to track the changes in vulnerability. The result shows that the level of vulnerability will be changed during the change of time period.

CHAPTER 8

LIMITATIONS AND FURTHRE STUDY

8.1 Further Studies

OSN can be analyzed for the following:

a) This work was only based on one status id. So, increase more status id to check the vulnerability of a user.

b) To study the behavior of a user.

c) To track the changes of vulnerability for short/long time.

d) To identify a node, whose role is more for vulnerability.

8.2 Limitations

- There must be the authentication between users through any Facebook application.
- There is a problem of UNIX time while accessing the friends' information as well as their friend list.
- For the Unix time expire problem can be solved using Facebook application
- User must be online or be the part of your Facebook application.
- Session key must be generated.
- Infinite session key will help to access the permission permanently.

Appendix

(Code for Implementation)

PHP CODING FOR FACEBOOK APPLICATION

```
<?php
require 'facebook.php';
// Create our Application instance.
// Get your application id and secret from Facebook to fill these in
$facebook = new Facebook(array(
  'appId' => '230856917016981',
  'secret' => '560aa22327d30b899dea98a4b637db56',
  'cookie' => true,
));
$fbUser = $facebook->getUser();
if ($fbUser) {
  try {
  $friends = $facebook->api('/me/friends'); // Get Friends
 } catch (FacebookApiException $e) {
  error_log($e);
  $fbUser = null;
  }
} else {
    $login_url = $facebook->getLoginUrl();
    header("Location: ".$login_url);
}
foreach($friends['data'] as $friend){
echo $friend['id'];
echo $friend['  '];
echo $friend['name'];
echo '<p></p>';
}
?>
```

Code Implementation for accessing friendlist through facebook

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using Facebook.Rest;
using Newtonsoft.Json.Linq;
using Facebook.Session;
namespace ConsoleApplication1
{
           class Program
          {
static string appKey = "230856917016981", sessionKey =
"2.AQBaolNr1808baBc.3600.1345960800.1-100002621473768", secrete =
"mVlaHWGFyOnUBRNVIc_ovw__", userid = "100002621473768", pageid = "",
postid = "100002621473768_251232768307437";
          static void Main(string[] args)
      {
         UpdateNotificationOnFB();
       }
       public static void UpdateNotificationOnFB()
       {
           Int32 likecounts = 0;
           Int32 commentcounts = 0;
           Api api = new Api(new DesktopSession(appKey,secrete,sessionKey));
           api.Fql.UseJson = true;
           try
        {
        var users = api.Fql.Query("SELECT user_id FROM like WHERE post_id='" +
postid  +"'");
       var userIdList = users.Replace("[", "").Replace("]", "").Replace("{\"user_id\":",
"").Replace("{", "").Replace("}", "").Split(',');
```

```
DataSet1TableAdapters.Tbl_Users_ListTableAdapter adapter = new
DataSet1TableAdapters.Tbl_Users_ListTableAdapter();
   foreach (var user in userIdList)
    {
     var name = api.Fql.Query("SELECT name FROM user WHERE uid='" + user
+ "'");
     name = name.Replace("[", "").Replace("]", "").Replace("{\"name\":",
"").Replace("{", "").Replace("}", "").Replace("\"","").Split(',')[0];
      adapter.Insert(user, name , "1");
    var friends = api.Fql.Query("SELECT uid2 FROM friend WHERE uid1='" +
user + "'");
     if (friends.Contains("error"))
        {
         continue;
        }
    var friendList = friends.Replace("[", "").Replace("]", "").Replace("{\"uid2\":",
"").Replace("{", "").Replace("}", "").Replace("\"", "").Split(',');
    foreach (var friend in friendList)
     {
   var frnname = api.Fql.Query("SELECT name FROM user WHERE uid='" +
friend + "'");
     if (frnname.Contains("error"))
      {
       continue;
      }
     var friendName = frnname.Replace("[", "").Replace("]",
"").Replace("{\"name\":",
          "").Replace("{", "").Replace("}", "").Replace("\"", "").Split(',')[0];
    adapter.Insert(friend, friendName, user);
     }
   }
  }
  catch
    {
            }
        }
        }
    }
```

References

[1] Alim S., Abdul-Rahman R., Neagu D. and Ridley M., "*Algorithm for data Retrieval from online Social Network Graphs*", IEEE, Conference on Computer and Information Technology (CIT 2010), in Proceedings of the 10th International IEEE Conference on Computer and Information Technology, Bradford, UK, 2010, pp 1660-1666.

[2] Alim S., Abdul-Rahman R., Neagu D. and Ridley M., "*Data Retrieval from Online Social Networking Profiles for Social Engineering Applications*", IEEE, International Conference for Internet Technology and Secured Transactions (ICITST-2009), in Proceedings of International Conference for Internet Technology and Secured Transactions, London, UK, 2009, pp 207-211

[3] Alim S., Abdul-Rahman R., Neagu D. and Ridley M., "*Online social network profile data extraction for vulnerability analysis*", J. Internet Technology and Secured Transactions, Vol.3,No.2, 2011, pp.194-209.

[4] Benevenuto F., Rodrigues T., Cha M. and Almeida V., "*Characterizing user behavior in online social networks*", Internet Measurement Conference : ACM, 2009.

[5] Catanese S., De Meo P., Ferrara E., Fiumara G. and Provetti A., "*Crawling Facebook for social network analysis purposes*", International Conference on Web Intelligence, Mining and Semantics, 2011.

[6] Catanese S., De Meo P., Ferrara E. and Fiumara G, "*Analyzing the Facebook Friendship Graph*", 2011.

[7] Chang C., Hsu C. and Lui S., "*Automatic information extraction from semi-structured web pages by pattern discovery*", Decision Support Systems - Web retrieval and mining, Volume 35 Issue 1, 2003

[8] Facebook, " *Press Room*", [Online] available from *http://www.facebook.com/press/info.php?statistics*, last accessed 28th January 2012.

[9] Ferrara E. and Baumgartner R., "*Automatic Wrapper Adaption by Tree Edit Distance Matching*", Proceedings of the 2nd International Workshop on Combinations of Intelligent Methods and Applications, March, 2011.

[10] Ghosh S., Korlam G. and Ganguly N., "*The effects of restrictions on number of connections in OSNs",* a case-study on twitter, 2010.

[11] Hansen D. and Shneiderman B., "*Analyzing social media networks: Learning by doing with NodeXL"*, Draft Version (7/07/09) with NodeXL Version 1.0.1.88.

[12] Hasib A., "*Threats of Online Social Networks*", IJCSNS International Journal of Computer Science, 2008

[13] http://en.wikipedia.org/wiki/List_of_social_networking_websites

[14] http://upload.wikimedia.org/wikipedia/commons/5/56/Facebook_popularity.PNG

[15] Laorden C., Sanz B., Alvarez G. and Bringas P.G., "*A Threat Model Approach to Threats and Vulnerabilities in On-line Social Networks*", Proceedings of the 3rd International Conference on Computational Intelligence in Security for Information Systems, 2010.

[16] Li X., Guo L. and Zhao Y. E., "*Tag Based Social Interest Discovery*", Proceedings of the 17th international conference on World Wide Web, 2008.

[17] Liben-Nowell D., "*An Algorithmic Approach to Social Networks*", Ph.D. Thesis, Massachusetts Institute of Technology, 2005

[18] Mohtasebi A. and Borazjani P. N., "*Privacy Concerns in Social Networks and Online Communities*", VALA session 14, 2010.

[19] Schank T., Wagner D., "*Approximating Clustering Coefficient and Transitivity*", Journal of Graph Algorithm and Applications, Vol 9, No 2, pp. 265-275, 2005.

[20] Tyagi N. and Gupta D., "*A Novel Architecture for Domain Specific Parallel Crawler*", Indian Journal of Computer Science and Engineering, Vol 1 no 1 44-53

[21] Wang D., Wen Z., Tong H., Lin C., Song C. and Barabási A., "*Information Spreading in Context*", 2011 in Proc. WWW, **2011**, pp.735-744

[22] Watts D.J. and Strogatz S.H., "*Collective dynamics of small world networks*", Nature, Vol. 393, No. 6684, pp.409–410, 1998

[23] Ye S., Lang J. and Wu F., "*Crawling Online Social Graphs*", Web Conference (APWEB), 12th International Asia-Pacific, 2010.

[24] Zeinalipour-Yazti D., Pallis G. and Dikaiakos M. D., *Chapter 8,* "*Online Social Networks: Status and Trends, New Directions in Web Data Management 1*", SCI 331, pp. 213-234