EVALUATION OF INFLUENTIAL NODES IN SOCIAL NETWORK

A DISSERTATION

Submitted in partial fulfillment of the requirements for the award of the degree
of
MASTER OF TECHNOLOGY
in
INFORMATION TECHNOLOGY

By

MUDUKURU DILEEP KUMAR

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE -247 667 (INDIA)
JUNE, 2011

CANDIDATE DECLARATION

I hereby declare that the work being presented in the dissertation report titled "**EVALUATION OF INFLUENTIAL NODES IN SOCIAL NETWORK**" in partial fulfillment of the requirement for the award of the **MASTER OF TECHNOLOGY** in **INFORMATION TECHNOLOGY**, submitted in the Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, is an authentic record of my own work carried out during the period from July 2010 to June 2011, under the guidance of Dr. Rajdeep Niyogi, in the Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee.

The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other Institute.

Dated:

Place: IIT Roorkee.

Mudukuru Dileep Kumar

CERTIFICATE

This is to certify that above statements made by the candidate are correct to the best ofour knowledge and belief.

Dated:

Place: IIT Roorkee.

Dr.RajdeepNiyogi

Assistant Professor,

Department of Electronics,

and Computer Engineering

ACKNOWLEDGEMENTS

First of all and foremost, I would like to express my deep sense of gratitude and indebtedness to my guide Dr. Rajdeep Niyogi, for his invaluable guidance and constant encouragement throughout the dissertation. His zeal for getting the best out of his students helped me to perform above my par. I am grateful to him for his help with the formalization and validation of the algorithm.

I would also like to acknowledge Institute Computer Center staff, IIT Roorkee for permitting me to use Computer Center their resources that have contributed to this research.

Mudukuru Dileep Kumar

ABSTRACT

Social network analysis has recently gained a lot of interest because of the advent and the increasing popularity of social media, such as blogs, social networks, micro-blogging, or customer review sites etc. Such media often serve as platforms for information dissemination, improving the performance of web search, recommendations in collaborative filtering systems, spreading a technology in the market. It is well known that the interpersonal relationships (or ties or links) between individuals cause change or improvement in the social system because the decisions made by individuals are influenced heavily by the behavior of their neighbors.

An interesting and key problem in social networks analysis is to discover the most influential nodes in the social network which can influence other nodes in the social network in a strong and deep way. 'Node position' is measure to evaluate the influence in social network. It enables to estimate how valuable the particular individual within the community is.

Some time we need to give high priority to the recent activities, but the previous algorithm fails to consider time (recent and old) in activities for evaluation of influential nodes. In this thesis proposed an enhancement which will solve this problem. New algorithm is compared with previous algorithm for its performance. We also proposed a solution to Top-k nodes selection problem. In Top-k nodes problem we need to select 'k' nodes which will maximize the information dissemination. To solve the problem combination of influence values calculated and new algorithm proposed are employed.

Table of Contents

LIST OF FIGURES AND TABLES

List of Figures

Figure No. **Page No.**

List of Tables

CHAPTER 1
INTRODUCITON

1.1 Introduction of social network

A social network is a social structure between nodes (people, groups, organizations, humans, animals, etc). It indicates the ways in which they are connected through various social relationships ranging from casual acquaintance to close familiar bonds.

After Internet boom, Internet has spawned many information sharing networks, the most well-known of which is the World Wide Web. Since a decade, a new class of information networks called "online social networks" has exploded in popularity and now rivals the traditional Web in terms of usage. Since nodes make relationship with other nodes through Internet, It is known as "online social network"[2]. Social networking sites such as MySpace (over 98 million users), Facebook (over 600 million users), Orkut (over 100 million users), and LinkedIn (over 70 million "professionals") 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.

Social network analysis (SNA) is the mapping of relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. The technique has been used rather effectively in the past and its applications continue to propagate. Social network analysis can be applied in academic researchers, business, law enforcement, and national security organizations to identify important connections and increase efficiency. Businesses have used social network analysis to analyze email patterns to determine which employees are overloaded. Law enforcement and national security agencies have exploited this technique to investigate users (nodes) in an organization and replicate the structure of networks. It may also be an effective tool for mass surveillance, for example the Total Information Awareness program was doing in-depth research on strategies to analyze social networks to determine whether or not U.S. citizens were political threats.

1.2 Motivation

Suppose that we are trying to market a product, or promote an idea, innovation or behavior, within a group (a social network). In order to do so, we have to "target" individuals for demonstrating an innovation, or explaining an idea (such as the consequences of alcohol use to teenagers). An important question is then whom we should target. Clearly, if there were no interaction between the individuals, this would be straightforward: the effect on each targeted individual could be determined in isolation, and we could choose the set of individuals with largest (expected) revenue or reach. However, individuals exit in form of complex social networks based on a mutual and multitude of different relations and interactions. Because of these interactions, they influence each other's decisions in adopting a product or behavior. These "target" individuals are known as "influential nodes".

One of the most meaningful and useful issue in social network analysis is evaluation of the influential nodes within the network. Since social network is complex network of humans, the problem of assessment the influential node becomes very complex because humans with their spontaneous and social behavior are hard predictable..However, the effort should be made to evaluate users who are the most influential among community members, possess the highest social statement and probably the highest level of trust [11]. These users can be representatives of the entire community.

In this thesis proposed a model that enhances previous approach on social networks, by involving time factor and overall influential users of a social network in the ranking process. My objective is to evaluate the influence indicating value (node position) of different users(nodes) in a social network. These values are used to find influential nodes in network. Please note that the same model can be applied to any social medium, for example blogs (where "influence" is considered the addition of a blog in one's blogroll), tweets (where "influence" is shown by following a tweeter), or consumer networks (where "influence" is shown explicitly by endorsement or reviews), or email communication network (where "influence" is shown by number of sent and received mails).

1.3 Problem statement

The aim of dissertation is to design and implementation of algorithm to find influential nodes using node position method in social network and compare its performance with other methods.

Objectives:

1. To model algorithm for computation of Node position value of nodes (influential nodes) in social network using node position method.
2. Enhance the algorithm by adding proposed method to include time factor in algorithm.
3. To design and implementation of algorithm to find top-k node from the influence values already calculated.
4. Comparing our algorithm with other centrality based algorithms and analysis of results.

1.4 Organization of Report

This report comprises of five chapters including this chapter that introduces the topic and states the problem. The rest of the dissertation report is organized as follows,

Chapter 2 gives an overview of Social Network Analysis. And discuss about the Data collection and processing models, properties. It also give brief literature review of the related work.

Chapter 3 gives details of proposed enhancement of the node position algorithm. Design, implementation and results obtained discussed here. Compared our algorithm with other centrality measures.

Chapter 4: gives details of algorithm to select the top-k nodes from the influential nodes. design and implementation and results are discussed here.

Chapter 5: Concludes the work and gives the directions for future work.

CHAPTER 2
BACKGROUND AND LITERATURE SURVEY

2.1 Background

Social network analysis is the mapping and measuring of relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. The nodes in the network are the people and groups, while the links show relationships or flows between the nodes.

After Internet boom people started communicating through email this is form the online social network. Online Social network is the social network maintained over Internet. Users join an online network, publish their profile and any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users.

An in-depth understanding of the graph structure of online social networks is necessary to evaluate current systems, to design future online social network based systems, and to understand the impact of online social networks on the Internet. For example, understanding the structure of online social networks might lead to algorithms that can detect trusted or influential users, much like the study of the Web graph led to the discovery of algorithms for finding authoritative sources in the Web [2].

There are many challenges are also involved in social network analysis. In this chapter of thesis a brief review of social network data collection methods, models to represent social network, properties of social networks are given.

2.2 Data collection process:

In most of Social network analysis cases only partial social networks are considered for analysis. There are several reasons for why partial social networks are used. First of all, social data is among the most valuable assets to the social network providers and is protected by privacy regulations/laws; therefore it is hard to get such data directly from the social network providers. Secondly, it is a great challenge for crawlers to collect millions of contact lists, profiles, pictures, videos, etc. from social networks. Many social networks use a large number of dynamic pages containing AJAX and DHTML effects, and it is not trivial to develop a parser to deal with such complex pages efficiently. Moreover, social network users are often provided with the flexibility to customize the layout of their pages, which further complicates the design and implementation of the parser. To make things worse, rate limiting is enforced by most social network, preventing crawlers from making many requests within a short period of time. Thirdly, as more users become concerned about their privacy in social network, many of them choose not to reveal their information to strangers, hence become "black holes" for data collection.

For data collection in Graphs model social network crawlers are used. With respect crawler perspective crawling large, complex graphs presents unique challenges.

2.2.1 Crawling connected component

The primary challenge in crawling large graphs is covering the entire giant connected component. Crawling the entire connected component is not feasible and one must resort it to small size using samples of the graph.

Sampling of graphs: three kinds of sampling method called node sampling, link sampling, and snowball sampling [6].

In *node sampling*, a certain number of nodes are randomly chosen and links among them are kept. The sampling fraction in this method is defined as the ratio of the number of chosen nodes (including isolated nodes that will be removed later) to that of all the nodes in the original network. Few isolated nodes generated due to no links to other nodes from selected list, so isolated nodes are neglected, so the number of nodes in a sampled network is a little bit less than that of selected nodes. In this sampling number of chosen links always depends on the number of nodes. Suppose the fraction of number of selected nodes is α and that of links among them is β. Then it is found that $\beta \sim \alpha^2$ if nodes pick randomly, since the maximum number of undirected links possible for n selected nodes are$\binom{n}{2} = \frac{n(n-1)}{2} \cong n^2$.

In *link sampling*, a certain number of links are randomly selected and nodes attached to them are kept. In *snowball* sampling, first choose a single node and all the nodes directly linked to it are picked. Then all the nodes connected to those picked in the last step are selected, and this process is continued until the desired numbers of nodes are sampled. To control the number of nodes in the sampled network, a necessary number of nodes are randomly chosen from the last layer. The snowball sampling method tends to pick hubs (nodes with many links)[6] in short step due to high connectivity of them. So whether the initial node is a hub or not does not make a noticeable difference in characterizing the sampled network.

The definition of snowball sampling, hubs are more likely to be selected by this method. Furthermore, once a hub is picked, every node connected to the hub is selected in the next step unless it belongs to the previous layer. This characteristic of snowball sampling tends to conserve the degrees of easily selected hubs, which leads to the decrease of degree exponents by holding the "tail" of the power-law degree distribution. So in social network graphs, collecting samples via the snowball method underestimate the power-law coefficient, but to more closely match other metrics, including the overall clustering coefficient. However, some previous studies of social networks have used small graph samples.

2.2.2 Common algorithms for crawling graph.

BFS(breadth-first search): Simply selecting the first item in the queue, breadth first search is probably the most popular one.

Greedy: The crawler selects the node with the largest degree in the queue. Since the nodes in the queue are not crawled yet, their degrees on the crawled subgraph are used.

Random: The crawler selects a node in the queue with probability proportional to its degree. This algorithm prefers nodes with large degrees, while also selecting nodes with small degrees to reduce sampling bias. Similar to the greedy algorithm, the degree here is computed on the crawled sub graph.

Hypothetical greedy: The crawler always selects the node with largest degree in the queue, while the degree here is the degree on the whole graph, i.e. the true degree. A typical application scenario for this algorithm is to sample a sub graph from a large graph.

Crawling directed graphs: Crawling directed graphs, as opposed to undirected graphs, presents additional challenges. Large graph have both Strongly Connected Component (SCC) and Weakly Connected Component (WCC). In SCC each node is connected with all other nodes in network, while in WCC it is not the same. Directed graph have forward links and backward links. Crawling only forward links does not necessarily crawl an entire WCC instead, it explores the connected component reachable from a set of seed users. This limitation is typical for studies that crawl online networks, such as the Web [5].

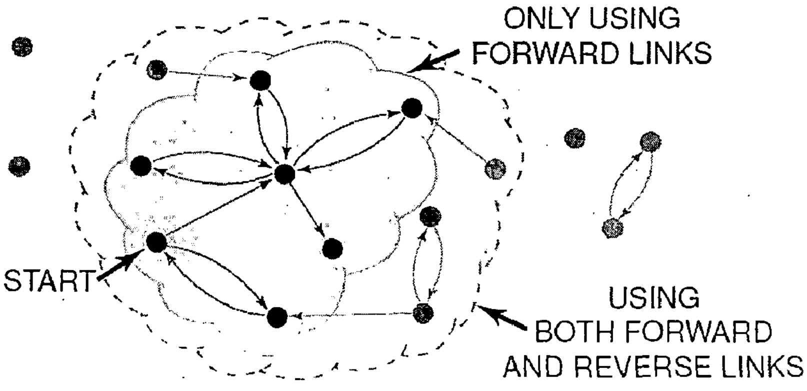

Fig 2.1 nodes reach by different link types

2.3 Social Network Models and Properties

After the data collection it has to be represented with efficient model before analyzing. There are many models are proposed for representing [8][20], among those some of them are discussed below. Graph based model is most popular model. It is well suited for representing node and relations.

Different social network models to represent are:

- Using Matrices to Represent Social network
- Statistical Models for Social Network Analysis
- Ontology Based Social.Network Analysis
- Using Graphs to Represent Social network

Using Matrices to Represent Social Relations:

The most common form of matrix in social network analysis is a very simple one composed of as many rows and columns as there are actors in our data set, and where the elements represent the ties between the actors. The simplest and most common matrix is binary. That is, if a relation is present, a one is entered in a cell; if there is no relation, a zero is entered. This kind of a matrix is the starting point for almost all network analysis, and is called an "adjacency matrix" because it represents who is next to, or adjacent to whom in the "social space" mapped by the relations that have measured. By convention, in a directed graph, the sender of a relation is the row and the target of the relation is the column. The representations to matrices outperform node-link diagrams(early method) for large graphs or dense graphs in several low-level reading tasks, except path finding.

Statistical Models for Social Network Analysis:

Statistical analysis of social networks spans over 60 years, Since the 1970s. Main idea of this method is to model probabilities of relational ties between interacting units (social actors), though only very small groups of actors were considered. Extensive introduction to earlier methods is provided by Wasserman and Faust [7]. Two of the most prominent current methods are Markov Random Fields (MRFs) introduced by Frank and Strauss and Exponential Random Graphical Models (ERGMs), also known as p*. The ERGM have been recently extended by Snijders et [2f] in order to achieve robustness in the estimated parameters.

Ontologies based Social Network analysis:

Ontologies are commonly deployed for the specification and explication of concepts and relationships related to a given domain. Social networks have the same purpose but with the focus on social relations and entities, hence domain ontologies related to social entities and relations can be designed and deployed. Through reasoning and inference ontologies do not allow the modeling of contradictory or inconsistent information. Modeling social networks via ontologies ensure the validity of the information encoded.

Graphs to represent social network: Social Networks is formalized as a graph $G := (V, E)$ which is an ordered pair of two sets. A set of vertices $V = (V_1, V_2, V_3, \ldots\ldots, V_n)$ which represents the social entity(node) and a set of edges $E = (E_{11}, E_{12}, E_{21}, ..., E_{ij})$ where E_{ij} represents the relation between the nodes i and j.

Properties of Social Network:

Social network properties there are some properties of social networks that are very important such as size, density, degree, reachability, distance, diameter, geodesic distance. Some more complicated properties which are used in social network analysis [1][6][9]. Few of them which are come across in this thesis are briefly explained below.

Maximal flow:

One notion of how totally connected two actors are, asks how many different actors in the neighborhood of a source lead to pathways to a target. If I need to get a message to you, and there is only one other person to whom I can send this for retransmission, my connection is weak - even if the person I send it to may have many ways of reaching you. If, on the other hand, there are four people to whom I can send my message, each of whom has one or more ways of retransmitting my message to you, then my connection is stronger. This "flow" approach suggests that the strength of my tie to you is no stronger than the weakest link in the chain of connections, where weakness means a lack of alternatives.

Power-law node degrees:

Power-law networks are networks where the probability that a node has degree k is proportional to $k^{-\gamma}$, for large k and $\gamma > 1$. The parameter γ is called the power-law coefficient. Researchers have shown that many real-world networks are power-law networks, including Internet topologies, the Web, neural networks, and power grids. The degree distributions of many complex networks, including offline social networks, have been shown to conform to power-laws [1]. Thus, it may not be surprising that social networks also exhibit power-law degree distributions. However some analysis shows, the degree distributions in social networks differ from that of other power-law networks in several ways.

Correlation of in degree and out degree

Studies of the in degree and out degree distributions in the graph helped researchers find better ways to find relevant information from graphs. Web is viewed as graph, the population of pages that are active i.e. have high out degree is not the same as the population of pages that are popular i.e. have high in degree [2]. For example, many Web pages of individual users actively point to a few popular pages like wikipedia.org or amazon.com. Web search techniques are very effective at separating a very small set of popular pages from a much larger set of active pages.

In social networks, the nodes with very high outdegree also tend to have very high indegree. Hence, active users i.e. those who create many links in social networks also tend to be popular i.e. they are the target of many links.

2.4 Measures in Social network analysis:

There are several methods used for social network analysis according to analysis purpose. Few of popular measures are node position [7], rank prestige , degree centrality [10][19] , closeness centrality[4], proximity prestige, betweenness centrality[20], and others.

Some of the measures used are explained below:

Betweenness centrality: Betweenness centrality BC of member x pinpoints to what extent x is between other members. Members with high BC are very important to the network because others actors can connect with each other only through them. It can be calculated only for undirected relationships by dividing the number of shortest geodesic distances (paths) from y to z by the number of shortest geodesic distances from y to z that pass through member x. This calculation is repeated for all pairs of members y and z, excluding x. Betweenness centrality of the member x is the sum of all the outcomes

$$BC(x) = \frac{\sum_{i \neq x \neq j, i, j \in m} b_{ij}(x)}{b_{ij}}$$

where:

b_{ij} (x) — the number of shortest paths from i to j that pass through x;

b_{ij} — the number of all shortest path between i and j;

m —the number of nodes in a network. If a member obtains high value of BC then it means that he/she is the node without which the network will split into subnetworks.

Indegree centrality: Indegree centrality(IDC), called also degree prestige, is based on the indegree number so it takes into account the number of members that are adjacent to a particular member of the community. In other words, more prominent people are those who received more nominations from members of the community.

$$IDC(x) = \frac{i(x)}{m-1}$$

where:

i(x) — is the number of members from the first level neighborhoodthat are adjacent to x;

m — the total number of members in the social network

Centrality Based on Node Degree: Out degree centrality (ODC) of the member x takes into account the number of out degree of the member x for edges which are directed to the given node.

$$ODC(x) = \frac{o(x)}{m-1}$$

where:

o(x) — the number of the first level neighbours to whom x is adjacent

m — the total number of members in the social network.

Users who communicate with the greater number of people obtain the greater outdegree centrality value. Actors with high out degree are recognized by other network members as a crucial cog that occupies a central location in a network. On the other hand users who have low outdegree centrality are not very open to the external world and do not communicate with many members.

Proximity Prestige: Proximity prestige PP(x) reflects how close all other members within the social community are to member x [37]. This measure depends on the geodesic distances, e.g. the length of the shortest paths d(x, y) from all members y to the considered member x.

$$PP(x) = \frac{\frac{p(x)}{m-1}}{\frac{1}{p(x)}\sum_{i=1}^{p(x)} d(x, y_i)} = \frac{p(x)^2}{(m-1)\cdot\sum_{i=1}^{p(x)} d(x, y_i)}$$

where:

p(x) — the number of all members y_i in the network who can reach member x, i.e. there exists a path from these members y_i to the given member x;

m — the number of nodes in a network

Closeness Centrality: The closeness centrality, in contrast to proximity prestige, pinpoints how close a member is to all the others within the social network. Its main idea is that the member takes the central position if they can quickly contact other members in the network. This measure emphasizes quality (position in a network) rather than quantity (number of links, like in a centrality degree measure). The member with high CC is a good propagator of ideas and information. A similar idea was studied for hypertext systems. The closeness centrality CC(x) of member x tightly depends on the geodesic distance, i.e. the shortest paths from member x to all other people in the social network and is calculated as follows

$$CC(x) = \frac{m-1}{\sum_{y \neq x, y \in M} c(x, y)}$$

Where:

c(x, y) — a function describing the distance between nodes x and y (i.e. max, min, mean or median);

m — the number of nodes in a network

Node Position method:

Node position (NP) is one of the measure to evaluate the position[7]. It enables to estimate how valuable the particular individual within the human community is. In other words, the importance of every member can be assessed by calculating their node position. Note that in this thesis the term Node Position is a metric used give influential indicating value to the node. In general, the greater node position one possesses the more valuable this member is for the entire community. It is often the case that we only need to extract the highly important persons, i.e. with the greatest node position. Such people are likely to have the biggest influence on others. As a result, we can focus our activities like advertising or target marketing solely on them and expect them to entail their acquaintances.

Let us consider the weighted social network *SN*(*M)* where *M* is the set of network members. The importance of member $x \in M$ in *SN*(*M)*, is expressed by the node position function, tightly depends on the strength of the relationships that this individual maintains as well as on the node positions of their acquaintances, i.e. the first level neighbors. In other words,

The member's node position is inherited from others but the level of inheritance depends on the activity of the members directed to the considered person, i.e. the intensity of common interaction, cooperation or communication. The activity contribution of one user absorbed by another is called commitment function. Node position *NP*(*x*) of individual *x* respects the values of node positions of the direct *x*'s acquaintances as well as their activities in relation to *x*. This NP(x) tell how much important is member with the social network SN(M).

Node position function *NP*(*x*) of individual *x* in the social network *SN(M)* considers the values of node positions of direct member's *x* acquaintances as well as their activities in relation to *x*:

$$NP(x) = (1 - \varepsilon) + \varepsilon \sum_{i=1}^{m_x} NP(y_i).C(y_i \to x)$$

where:

- y_i — *x*'s acquaintances, i.e. the members who are in direct relationship to *x*:
- $C(y_i \to x) > 0$;

- mx — the number of x's acquaintances.
- ε – the constant coefficient from the range [0; 1];
- $C(y_i \rightarrow x)$ – the function that denotes the contribution in activity of y_i directed to x.

Top-k problem: Define an influence function σ()as follows. If 'P' is the set of initially active nodes (also called the target set), then σ(P) is the expected number of active nodes at the end of the diffusion process. For economic reasons, we want to limit the size of the initial active set P . For a given constant , the top-k nodes problem seeks to find a subset of P nodes of cardinality that maximizes the value of σ(P).

The algorithmic and computational aspects of the top-k nodes problem are investigated by [25][12]. The authors showed it as optimization problem of selecting the most influential nodes is NP-hard and derive the first provable approximation guarantees for the proposed algorithm. The authors first showthat this objective function is a submodular function under the linear threshold model and the independent cascade model. A function g() is sub modular if it satisifies g(P U {i}- g(P))≥g(T U {i} – g(T) for all elements of i and all pairs of P subset of T. greedy algorithm achieves an approximation guarantee of (1-(1/e)) where $e=\sum_{r=1 to \infty}(1/r!)$.

2.5 Literature survey

As explained above one of the most meaningful and useful issue in analysis is the identification of the most "influential" nodes (users) in social network. Research done in this area are discussed in this section.

Research in [18] first proposed that targeting a few key individuals may lead to strong "word-of-mouth" effects, wherein friends recommend a product to their friends, who in turn recommend it to others, and so forth.

In [2] J. Kleinberg studies shown Influential users act as hubs within their community and thus play a key role in spreading information. This has obvious implications on "word of mouth" and product marketing, as indicated in study [3] by Domingos, P, Richardson, which in turn makes influential users important for the promotion and endorsement of new products or ideas.

In [8] given a model to identification of influencers as a combinatorial optimization problem: given a fixed number of nodes that can be initially activated or infected, find the set of nodes with maximum influence over the entire network - the one that generates the largest cascade of adoptions.

Researchers have investigated the identification of likely influential users through centrality analysis techniques out degree centrality, eccentrality, closeness centrality, or betweeness centrality [10][19][20],

Centrality is the concept of being "in the thick of things." In 1978 Freeman reviewed and clarified a growing field of research on centrality of nodes for binary networks in an article published in the first issue of Social Networks[22]. Three measures were formalized: degree, closeness, and betweenness. Degree was the number of ties or neighbours of a node; closeness was the inverse of the sum of all shortest paths to others or the smallest number of ties to go through to reach all others individually; and betweeness was the number of shortest paths on which a node was on.

The three measures have already been generalized to weighted networks. Barrat et al[23] generalized degree to weighted networks by taking the sum of weights instead of the number ties, while Newman[21] and Brandes [13] utilized Dijkstra's (1959) algorithm of shortest paths for generalizing closeness and betweenness to weighted networks, respectively. Dijkstra's algorithm defined the length of paths as the sum of cost which is generally only defined as the sum of the inversed tie weights. Many researches implemented parallel version of centrality measure but all these generalizations fail to take into account the main feature of the original measures formalized by Freeman[22] the number of ties.

Node position measure is proposed in [7] by Piotr Bródka, user activity to the network is consider and also number ties. It is based on page-ranking algorithm. In this thesis node position algorithm is enhanced and implemented and compared with other methods.

Recently W. Chen, Y. Wang, and S. Yang in [25] given time solution to maximize influence in the social network. They modified greedy model for influence measuring.

2.6 Research Gaps

The existing measures suffer from many shortcomings. Most important are listed below.

1. The methods mentioned previously not consider the "time factor" for determining influential nodes. Since in real world the importance of relation changes with time we need method which gives more importance to present relationship than old relationships.
2. Most of the previous methods are very inefficient when applying them in complex networks which constitute big part of the networks existing in the Internet. For example the calculation of the shortest paths within the large networks is a very time consuming task[9]. Even though Node position methods works in large graph, its taking more processing time.
3. *Disconnected Networks:* Methods which use shortest paths and eigenvector to measure cannot be applied in disconnected graphs that exist in the Internet. The calculation of these measures values for a disconnected graph gives as the outcome the values zero for each node. But node position method works in these disconnected networks.

CHAPTER 3
EVALUATION OF INFLUENCE VALUE OF NODES

Node position (NP) is one of the measures to evaluate the influence [7]. It enables to estimate how valuable the particular individual within the human community is. In other words, the importance of every member can be assessed by calculating their node position. Note that in this thesis the term 'Node Position' is a metric which indicates influence value of the node. In general, the greater node position one possesses the more valuable this member is for the entire community.

The reason behind choosing this algorithm is, it considers the strength of the relationship, of course the centrality measures include strength of relationship in calculation, and however the issue of assessing the connection strength is not considered. i.e unlike the centrality algorithm it takes both strength of the relationship and number of relationships into consideration.

Node position measure:

Node position function *NP*(*x*) of individual *x* in the social network *SN(M)* considers the values of node positions of direct member's *x* acquaintances as well as their activities in relation to *x*:

$$NP(x) = (1 - \varepsilon) + \varepsilon \sum_{i=1}^{m_x} NP(y_i).C(y_i \to x) \tag{3.1}$$

where:

- y_i — x's acquaintances, i.e. the members who are in direct relationship to x:
- $C(y_i \to x) > 0$;
- m_x — the number of x's acquaintances.
- ε – the constant coefficient from the range [0; 1];
- $C(y_i \to x)$ – the function that denotes the contribution in activity of y_i directed to x.

The value of ε denotes the openness of node position measure on external influences: how much x's node positions are more static and independent (small ε) or more influenced by others (greater ε). In other words, the greater values of ε enable the neighborhood of node x to influence the x's nodes position to a large extent.

In general, the greater node position one possesses the more valuable this member is for the entire community. It is often the case that we only need to extract the highly important persons, i.e. with the greatest node position. Such people are likely to have the biggest influence on others. As a result, we can focus our activities like advertising or target marketing solely on them and we would expect that they would entail their acquaintances. The node position of user x is inherited from the others but the level of inheritance depends on the activity of the users directed to this person, i.e. intensity of mutual communication. Thus, the node position depends both on the number and quality of relationships.

3.1 Commitment function

Social network is represented using Graph G(x,y). weight of the edge represent the relation. Weight of represents the attribute value of relation. For example In network if the edge represents the Email communication then weight tells us the how many Emails sent or received.

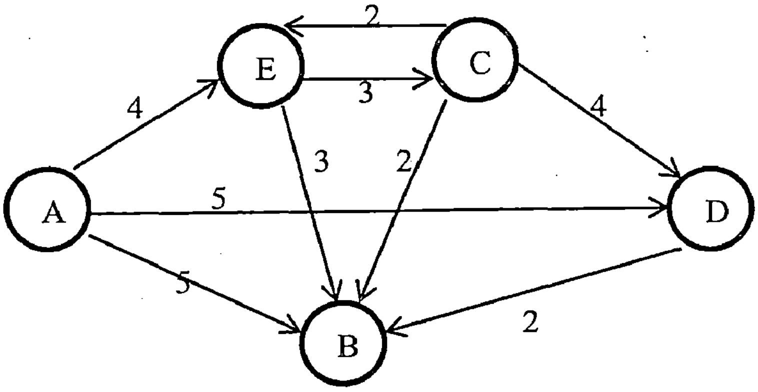

Fig 3.1 Example of Social Network

To assess the strength of the relationship between two individuals x and y within the virtual social network the commitment function C(x→y) is used. It denotes the amount of the member y's activity that person y passes to member x and is easily derived from relationship commitment function C(x→y).

The commitment $C(\text{x}\rightarrow y)$ of member y within activity of their acquaintance x is directly evaluated from source data as the normalized sum of all contacts, cooperation, and communications or any form of "influence"(mentioned in chapter1) from y to x in relation to all activities of y

$$c(x \rightarrow y) = \begin{cases} \frac{A(x \rightarrow y)}{\sum_{i=1}^{m} A(x \rightarrow yi)} & when \sum_{i=1}^{m} A(x \rightarrow yi) > 0 \\ 0 & when \sum_{i=1}^{m} A(x \rightarrow yi) = 0 \end{cases} \tag{3.2}$$

Where: $A(x \rightarrow y)$ – the function that denotes the activity of node x directed to node y, e.g. the number of emails sent by x to y; $A(x \rightarrow y) \geq 0$.

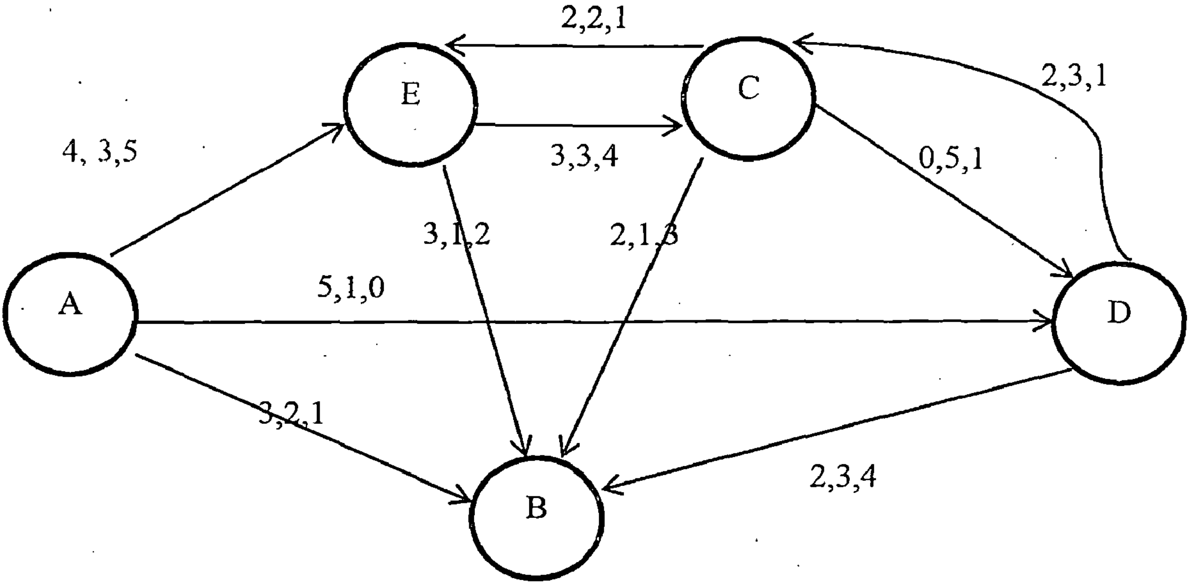

Fig 3.2 Example of time series activities in SN

m – The number of all nodes within the social network

But there is problem in above mentioned commit function. Some time we need to give high priority to the recent activities. Assume user(node) is very active for few months, after few months user started becoming inactive. Nodes will totally become in-active after few more months. i.e user commination in network is decreasing day by day. For example in Fig 4.2, Fig 4.2 shows the activities of nodes over a period of 3 months. Observer the A's activities (5, 1, 0), it is decreasing every month. If you apply commitment function to the user, it will give high "C" value to the user even though user is in-active, since he was very active in the past. But above mentioned function is failed to consider the change.

In this thesis I proposed a solution to above problem and solution is implemented and tested. The data collected from the online social network from specific duration. So we have the time of data with us. So I divided the data according to day or week or month or into specific intervals based on time. Activities are divided into k periods. Activities in each period are considered separately for each individual in commitment function.

$$c(x \rightarrow y) = \begin{cases} \frac{\sum_{j=1}^{k} \lambda^j A_i(x \rightarrow y)}{\sum_{i=1}^{m} \quad \sum_{j=1}^{k} \lambda^j A_i(x \rightarrow y)} & when \sum_{i=1}^{m} A_i(x \rightarrow y) > 0 \\ 0 & when \sum_{i=1}^{m} A_i(x \rightarrow y) = 0 \end{cases} \tag{3.3}$$

where: $A(x \rightarrow y)$ – the function that denotes the activity of node x directed to node y, e.g. the number of emails sent by x to y; $A(x \rightarrow y) \geq 0$.

m – the number of all nodes within the social network

λ- is the constant co-efficient which ranges from 0 to 1 (which is used to tell how much priority is given to old records and new records)

Note: $A(x \rightarrow y) = 0$ i.e emails sent to themselves are excluded.

The activity of person y is calculated in every time period and after that the appropriate weights are assigned to the particular time periods, using λ^i factor. The most recent period $\lambda^i = \lambda^0 = 1$, for the previous one $\lambda^i = \lambda^1 = \lambda$ is not greater than 1, and for the most former period $\lambda^i = \lambda^{k-1}$ receives the smallest value. For example, if one year's data set is proceeded and a period is a month then k=12. For λ=0.7, the data from January is considered with the factor 0.7^{11}=0.019, for February we have 0.7^{10}=0.0285, ..., for October 0.7^2=0.49, for November – 0.7 and finally for December 0.7^0=1. This in a sense is similar to an idea which was used in the personalized systems to weaken older activities of recent users.

One of the activity types is the communication via emails. In this case, $A_i(x \rightarrow y)$ is the number of emails that are common for x and y in the particular period i; and $\sum A_i(x \rightarrow y)$ all chats in which x took part in the ith period. If person x had many common chats with y in comparison to the number of all x's chats, then y has greater commitment within activities of x, i.e. C(x→y) will have greater value and in consequence the social position of member y will grow.

The time complexity of calculating commitment function is O(mn) where m- total no.of number elements in the network, n-no.of acquaintance. Since social network exhibits the powerlaw and small world phenomenon n<<m. so mn<<m^2. So complexity is linear.

The value of commitment function $C(y \rightarrow x)$ in the social network must satisfy the following set of criteria:

1. The value of commitment is from the range [0; 1]: $\forall(x, y \in Nodes)\ C(x\rightarrow y) \in [0; 1]$.
2. The sum of all commitments has to equal 1, separately for each node of the network: $\forall(x \in Nodes) \sum x \in Nodes\ C(x\rightarrow y) = 1$
3. If there is no relationship from y to x then $C(x\rightarrow y) = 0$
4. If a member y is not active to anybody and other n members x_i, $i = 1,...,n$ are active to y, then in order to satisfy criterion 2, the sum 1 is distributed equally among all the y's acquaintances x_i, i.e. $\forall(x_i \in Nodes)C(y \rightarrow x_i) = \frac{1}{n}$

Since the relationships are not reflexive and with respect to criterion 3, the commitment function to itself equals 0.

According to the above criteria all values of commitment are from the range [0, 1] (criterion 1) as well as the sum of all commitments equals 1, separately for each member of the network (criterion 2). Moreover, there is no relationship x to y so $C(x \rightarrow y) = 0$ (criterion 3). In Fig 4.1 B is not active any node but all node are active with the B so according to condition 4, the commitment of B is equally distributed among all B's connections

$$C(B\rightarrow A)=C(B\rightarrow C)= C(B\rightarrow D)= C(B\rightarrow E)= 1/4.$$

Algorithm 1: Commitment Evaluation Algorithm

Input: D- containing data about the communication activities between members in social network with M nodes

Output: C- list that stores commitment value for each order pair of nodes

```
begin
intialise all members to zero
commitment_of_x:=0;
acquaintances_of_x:=0;
for (each member $y \in M$) do
begin
Crel[x,y]= weight of edge (x,y)*$\lambda_i$)/(totaledgeweight);
commitment_of_x:=commitment_of_x+Crel[x,y];//
if (Crel[y,x]>0) then
acquaintances_of_x:=acquaintances_of_x+1;
end;
for (each member $y \in M$) do
    if (Crel[x,y]>0) then
        C[x,y]=Crel[x,y];
   else
        if (commitment_of_x=0 and Crel[y,x]>0) then
        C[x,y]=1/acquaintances_of_x;
      else
        C[x,y]=0;
end;
end.
```

3.2 Node Position Calculation:

The node position is calculated in the iterative way, i.e. the left side of Eq. 1 is the result of iteration while the right side is the input:

$$NP_{n+1}(x) = (1 - \varepsilon) + \varepsilon \sum_{y \epsilon M} NP_n(y_m).C(y \to x) \qquad (3.4)$$

where: $NP_{n+1}(x)$ and $NP_n(x)$ — the node position of member x after the n + 1st and nth iteration, respectively. To perform the first iteration, we also need to have an initial value of node position $NP0(x)$ for all $x \in M$:

$$NP_1(x) = (1 - \varepsilon) + \varepsilon \sum_{y \epsilon M} NP_0(y_m).C(y \to x) \qquad (3.5)$$

Since the calculations are iterative, we also need to introduce a stop condition. For this purpose, a fixed precision coefficient τ is used. Thus, the calculation is stopped when the following criterion is met:

$$\forall(x \epsilon M)|NP_n(x)| \leq \tau \qquad (3.6)$$

Obviously, another version of the stop condition can be also applied, e.g.: $|SNP_n - SNP_{n-1}| \leq \tau$

where: SNP_n and SNP_{n-1} — the sum of all node positions after the nth and nth iteration, respectively. Based on Eq. 3.1 the NP algorithm (Position In the Network) was developed.

For algorithm NP name is given since its calculates the Node position of all node from node perspective, i.e the node position is calculated one by one for each network node. First, two lists *NPprev* and *NPnext* that contain the node position values are created. *NPprev* serves to store social positions from the previous iteration whereas the node positions calculated in the current iteration are stored in *NPnext*.

At the beginning, the initial node positions values are assigned to the elements from *NPprev*. Afterwards, for each member x from M its *NPnext* is set to 1 $-\varepsilon$. Next, for each member y from M the value of commitment function $C(y \to x)$ is multiplied by *NPprev*[y] and by ε.

The result of this operation is added to the current value of x's social position that is stored in *NPnext*[x]. Finally, the values from *NPnext* are assigned to *NPprev* and the iteration is finished. The next iteration is performed unless the stop condition is met. Complexity of algorithm is O(nm) where m is total number of nodes, n is the maximum number of acquaintances (connectednode);

NP algorithm

Input:

C – List Containing commitments for each ordered pair $(x1, x2) \in M$,
NP-Initial Node positions of *m nodes*
ε - coefficient from Equation 1, $\in [0; 1]$,
τ - stop condition

Output:

NP- final node positions,

Node with highest node position

```
begin
$NP_{prev} := NP0$
repeat
begin
 for( each member x from M do
  begin
     $NP_{next}[x] := (1 - \varepsilon)$;
     for( each member y from M do
          $NP_{next}[x] := NP_{next}[x] + \varepsilon \cdot NP_{prev}[y] \cdot C[y, x]$;
     end;
     $NP_{prev} := NP_{next}$;
end;
until stop condition 5 is fulfilled for all members;
Return node with highest NP
end.
```

3.3 Implementation:

In this thesis algorithm is implemented in C++ language. Data set is downloaded from http://toreopsahl.com/datasets/ It is free for students doing research and it can accessed free. It is the data about an online Facebook like network among students at University of California, Irvine. The edge list includes the users that sent or received at least one message during that period (1,899). A total number of 59,835 online messages were sent among these over 20,297 directed ties [24]. And also time series of same data set is also used. In time series data set the messages sent were a divide in period of months.

Data set is fetched from file and is stored in array of adjacency list for processing. For every edge(x,y) x will give array index and y is stored as linked list to it along with its weight. Commitment function takes adjacency list of an each node as argument and calculates the commitment value for each node. Criteria 4 for commitment function is satisfied or not checked for all nodes. These commitment values are stored in the structure comt.

Data structure used are store graph and commit value

```
Struct Associate
{
      Int total;
      Int  source;
      List *link;
};

Sturct List{
      Int dest;
      Float commit;
      Float weight;
      List *link;
};

Struct List{
     Int dest;
     Float commit;
     Float w1,w2,w3;
     List *link;
};
```

```
Struct Comt{
             Int sr;
             cList *link;
};
Sturct cList{
             Int dest;
             Float comt;
             cList *link;
};
```

3.4 Results

The Test bed: 2.4GHz core2duo processor. 4GB RAM.

3.4.1 Comparison of Node position algorithms

Execution time of the algorithm without considering Time Factor : 3.465 s

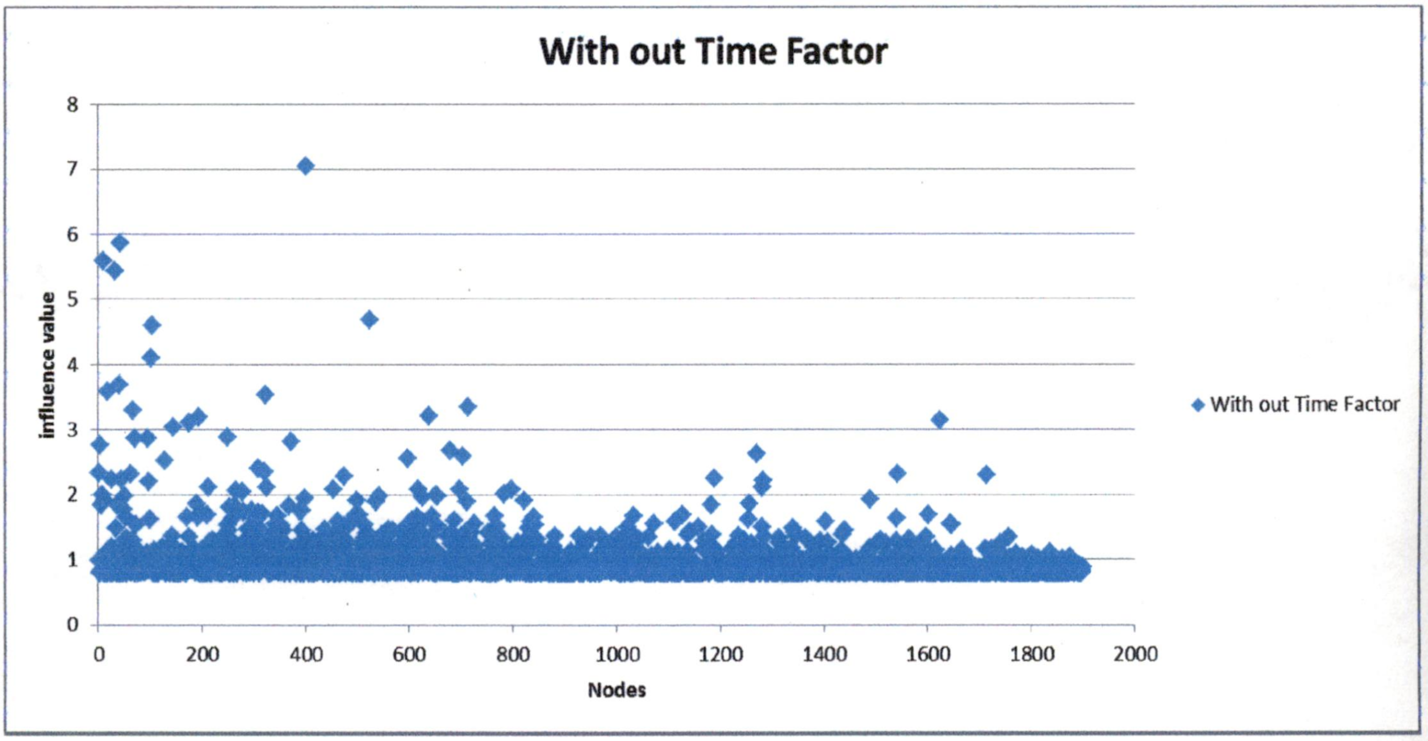

Fig 3.3 : Graph showing node and influence computed without considering Time factor

Execution time of the algorithm with considering Time Factor : 3.509 s

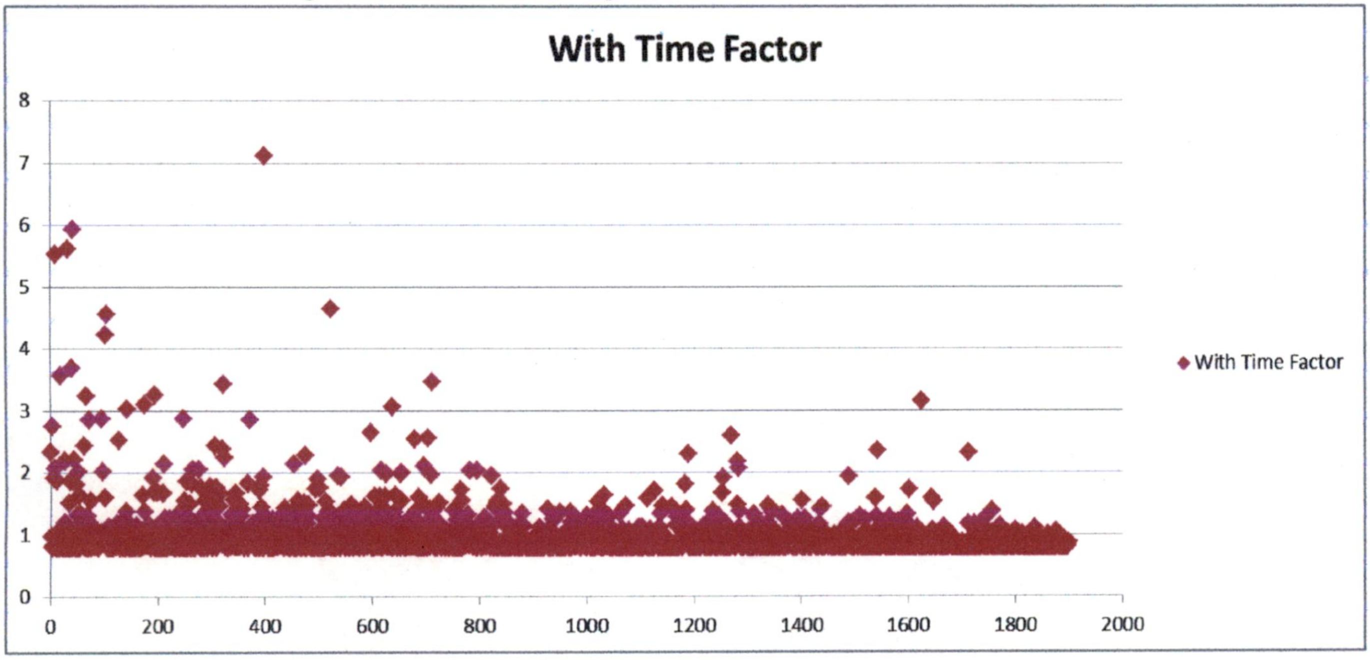

Fig 3.4 : Graph showing node and influence computed with considering Time factor

Algorithm 1 (with time factor)		Previous algorithm (with out time factor)	
Node	Influence value	Node	Influence value
400	7.10967	400	7.04306
42	5.9256	42	5.8708
32	5.61249	9	5.59599
9	5.5329	32	5.44361
523	4.65042	523	4.67733
105	4.56327	105	4.60209
103	4.22202	103	4.10255
41	3.69835	41	3.69453
19	3.57389	19	3.58618
713	3.45894	323	3.53778
323	3.4325	713	3.34995
194	3.26359	67	3.30192
67	3.24084	638	3.21003
1624	3.15139	194	3.20018
176	3.10278	1624	3.1377

Table 3.1 comparison of 'algorithm 1' with previous algorithm

Fig 3.3 shows the nodes and its influence values derived from the node position algorithm. In this Time series data set is not used. Fig 3.4 shows the nodes and their influence values derived from the modified algorithm. In this Time series data set is used. Both version execution times are of difference 0.05seconds. Table 3.1 compares two algorithm results note that in this only 15 out 1899 nodes results consider.

Observe Table 3.1. let's call with time factor column as "new" results and without time factor results as "Old" results. By observation we can see the changes in influence of all the values. But there are only few nodes shown visible variance in results. Consider node 32, in old result its value is less than node 9 value. But in New result node 9 influence in increase. Similarly node 194 influence value is increased. That means these are the node becoming active with time. Maybe after few more months it will become top influence node if present situation continues.

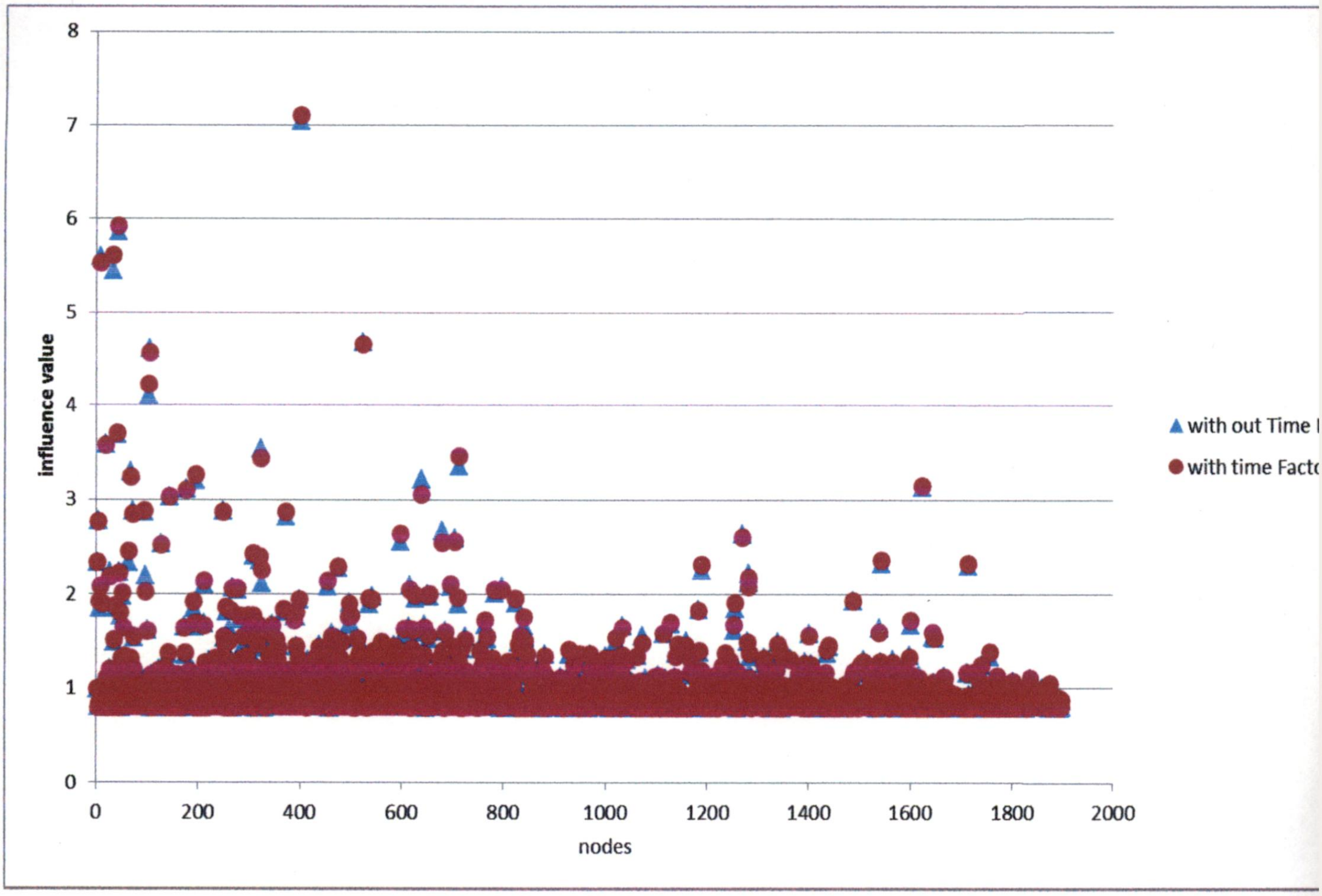

Fig 3.5 : comparison of algorithm 1 with previous algorithm

Fig 3.3 gives the Comparison of change influential nodes. The blue ‘triangles’ are indicates Old results (without time factor) and red ‘circles’ are New results. Observe the table 3.1 nodes; we can see the visible shift in the influence value.

3.4.2 Comparison of Node position algorithm for different λ values

λ is the constant co-efficient which ranges from 0 to 1. Which used for including time factor in algorithm. Experiments are performed to find out which is the optimal value for λ in calculation. Experiments are performed for λ= 0.1 to 0.9. As see from Fig 3.6 there is not much difference in results. So we can use with 0>λ>1 for influence calculation.

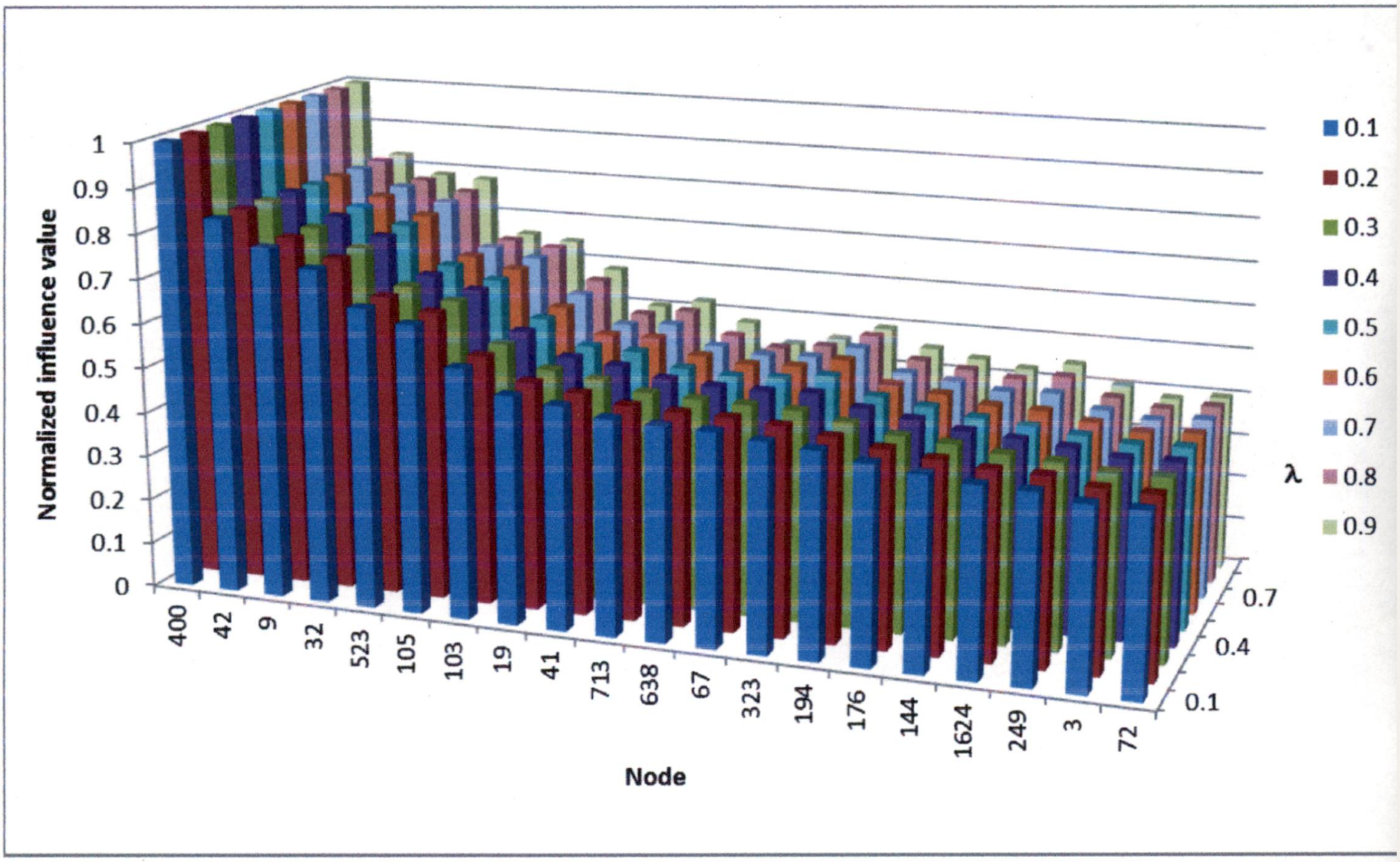

Fig 3.6 Comparision of algorithm for different λ values

3.4.3 Performance analysis of algorithm with other measures

We compared our algorithm with the Betweeness Centrality, Closeness Centrality, degree Centrality. These Centrality algorithms are implemented using YFiles Library [27]. YFiles works provides Library for analyzing of graphs. YFiles algorithms are the faster algorithms available in internet. In this section we are testing our algorithm for its fastness and its accuracy.

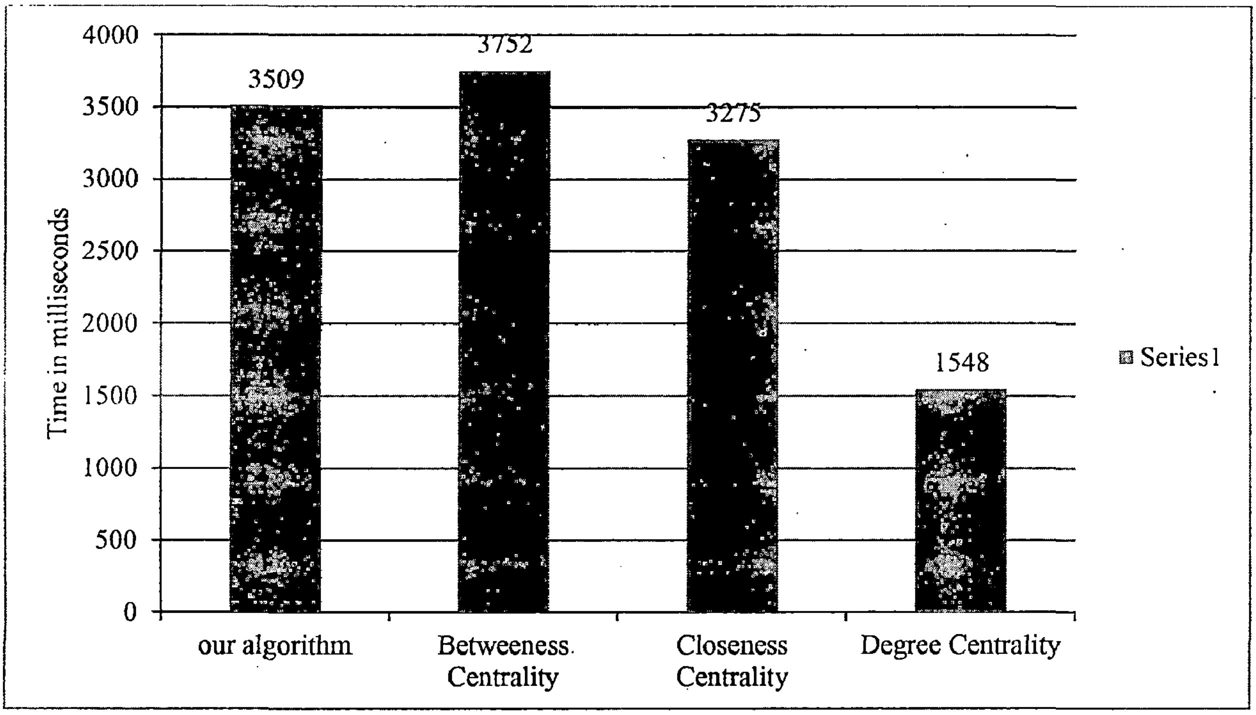

Fig 3.7 performance comparison of our algorithm with others

Degree was the number of relationships or neighbors of a node; closeness was the inverse of the sum of all shortest paths to others or the smallest number of relationship to go through to reach all others individually; and betweeness was the number of shortest paths on which a node was on. Betweeness centrality tell how important node with in network more accurately than closeness centrality. Degree centrality tell what is the strength followers.

In Fig 3.7 we can see that our algorithm is faster than the betweeness centrality but slower than closeness centrality and degree centrality. It is giving satisfying results than betweeness centrality.

In Fig 3.8 we can see that all the influential nodes calculated have the high Betweeness centrality and Closeness centrality measures also. So the influential nodes selected after algorithm have high centrality measures also.

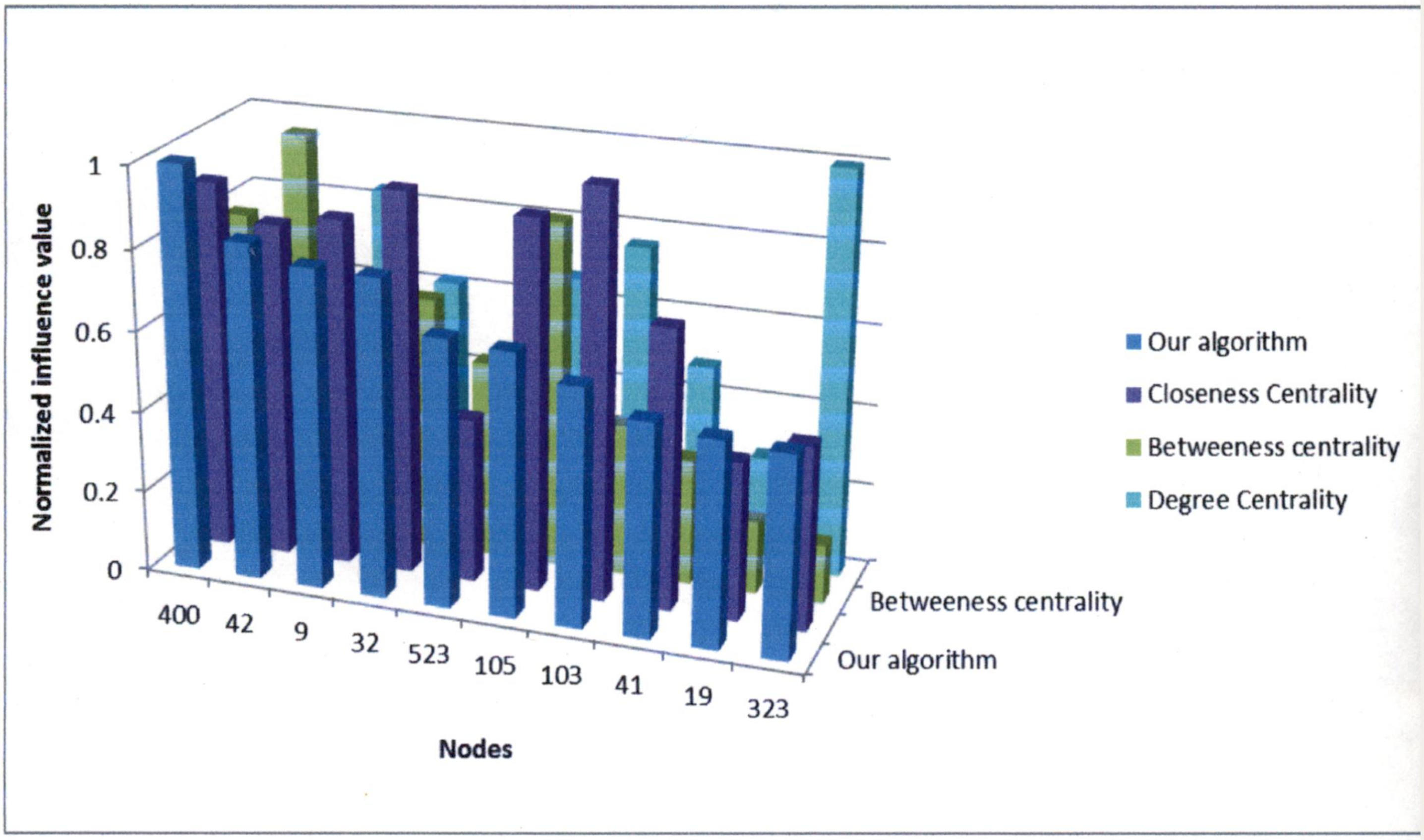

Fig 3.8 Comparison of nodes influence values of our algorithm with others centralities

CHAPTER 4
TOP-K NODES SELECTION

The node position evaluation algorithm evaluates the influence values based on its activeness in the community. Nodes with highest value are the most influential nodes in the network. This node position can be applied to any social medium.

4.1Top-k nodes problem:

Now, we come to the important question of how to choose the top-k nodes from the RankList. The naive approach is to choose the first k in the RankList as the top-k nodes. This approach suffers from the following drawback: the chosen nodes may be clustered at one place.

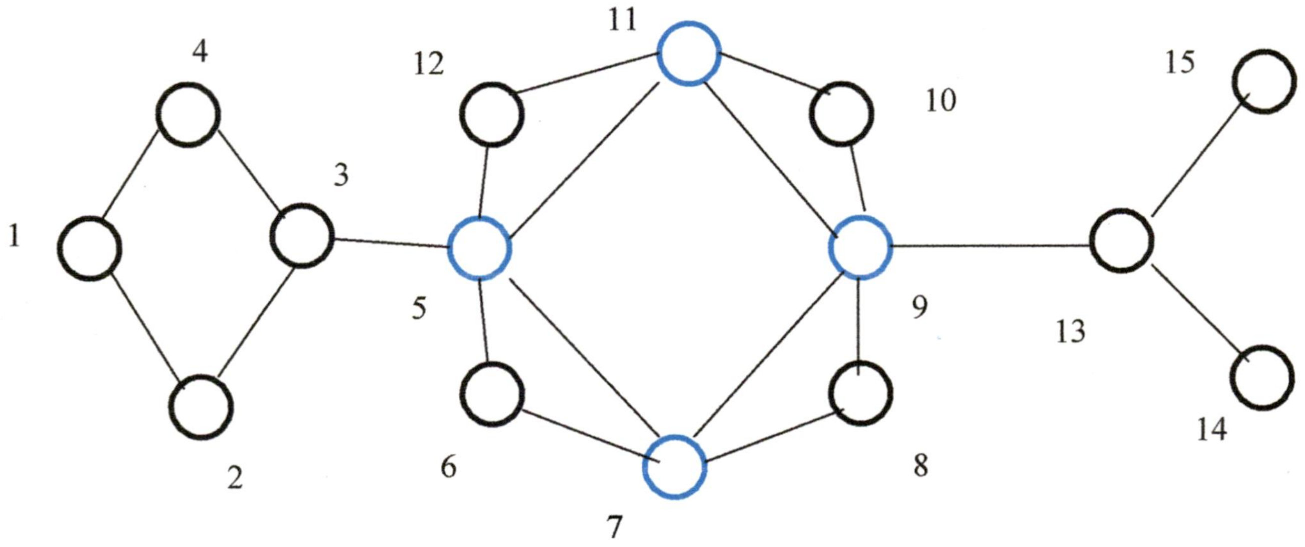

Fig 4.1 Example Network

For example, consider the network shown in Fig 4.1. assume that after applying node position algorithm we obtained RankList={5,7,9,11,3,13,12,4,10,2,6,8} if we want to choose 5 top influencing nodes then according to rank list 5 nodes are {5,7,9,11,4}. Observe that four out of five are located in the same cluster of the network. On the other hand, if these nodes are appropriately spread over the network, then such a situation will increase the information dissemination rate.

4.2 Algorithm proposed

Motivated by the above situation, to solve above problem the following algorithm is proposed:

Consider the nodes in the order given by RankList and initially we add the nodes to the list of top- nodes that are not adjacent. More precisely: First node is not compared with any nodes since it has the highest influence value or node position value. Second node from the RankList is compared with the node above it i.e first node. if second node is adjacent to first node then it is not selected for Top-K nodes. select the node to Top-K nodes only when it is not adjacent.

Social network is directional graph, so if a node A is connected node B doesn't mean B is connected A. there may be probability that B is having connection with node A. So we can't say two nodes adjacent only comparing one direction. Also if A is not connected B, B is not connected C, then we can't say A is not connected to C in social network. So node has to be compared with all other nodes in RankList for connectivity.

In this proposed algorithm comparisons with all other nodes is avoid, only comparision with node above it in RankList are performed. The complexity will be reduced to O(mnk). Where m is the number of nodes and n is the number of adjacent node of K nodes. Algorithm is:

i) we take the first node from the RankList and add it to the list of top- nodes;

ii) we take the second node from RankList and add it to the list of top- nodes if it is not adjacent to the node in the list of top-k nodes, and so on; Here two cases are considered

case 1: Is $(i+1)^{th}$ node is adjacent to all the nodes of adjacency list of i^{th} node

case 2: i^{th} is not adjacent to all the nodes of adjacency list of i^{th} node

iii) in general, when we consider a node from the RankList, we add it to the list of top- nodes if it is not adjacent to any node in the list of topnodes.

In this process, after adding a certain number of nodes to the list of top- nodes, we may not find any node from the RankList that is not adjacent to any node in the list of topnodes. In other words, any node in the network is either added to the list of top-k nodes or adjacent to some node in the list of top-k nodes. Then, we consider nodes with the highest node position values that are still not included to the list of top- nodes and add them to the list. We stop the above process when the size of the list of top-k nodes is k.

Algorithm 2

```
1   Input:
        a. Ranklist- set containing sorted nodes according their node position value
        b. Nodes adjacency list AdjList of all nodes
2   Ouput:
        a. TopList- containg nodes which maximize the influence
3   For each node in Ranklist[ i:1 to n ]
4               Case1: for each node in AdjList[1to m] of $i^{th}$ node
5                           If  $(i+1)^{th}$ node is found in adjacency list
6                               RankList=RankList-RankList[i+1];
7                                   Go to End1;
8               Case 2: for each node in AdjList[1to m] of $(i+1)^{th}$ node
9                           If  $i^{th}$ node is found in adjacency list
10                              RankList=RankList – $(i+1)^{th}$ node
11              If node is not found in above case1 and case2
12                      Add node to TopList
13                      TopList=TopList+$(i+1)^{th}$ node
14          Repeat steps 4 to 12 for each node above TopList
```

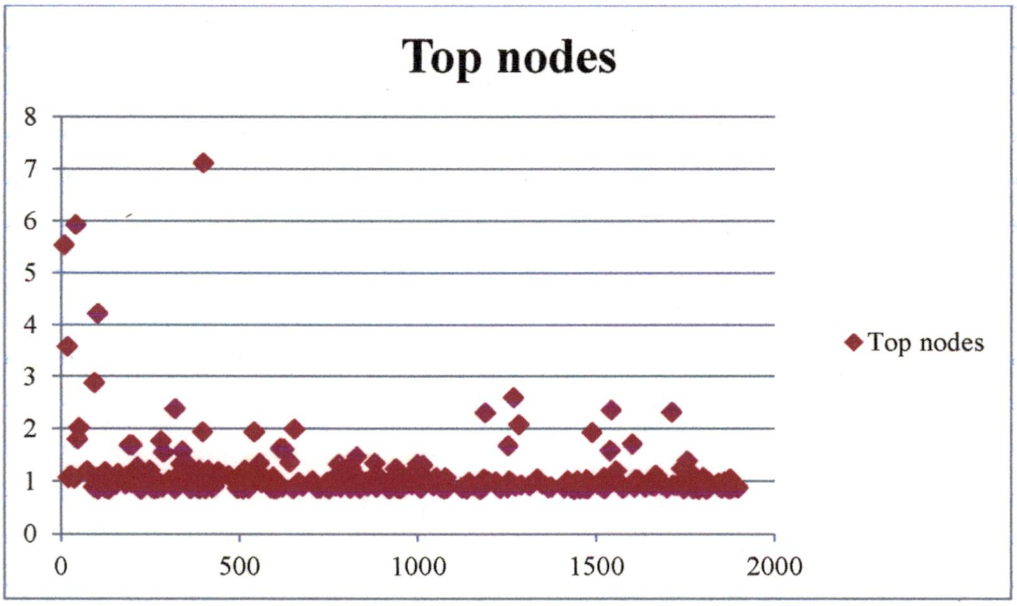

Fig 4.3 Nodes Filtered after algorithm 2

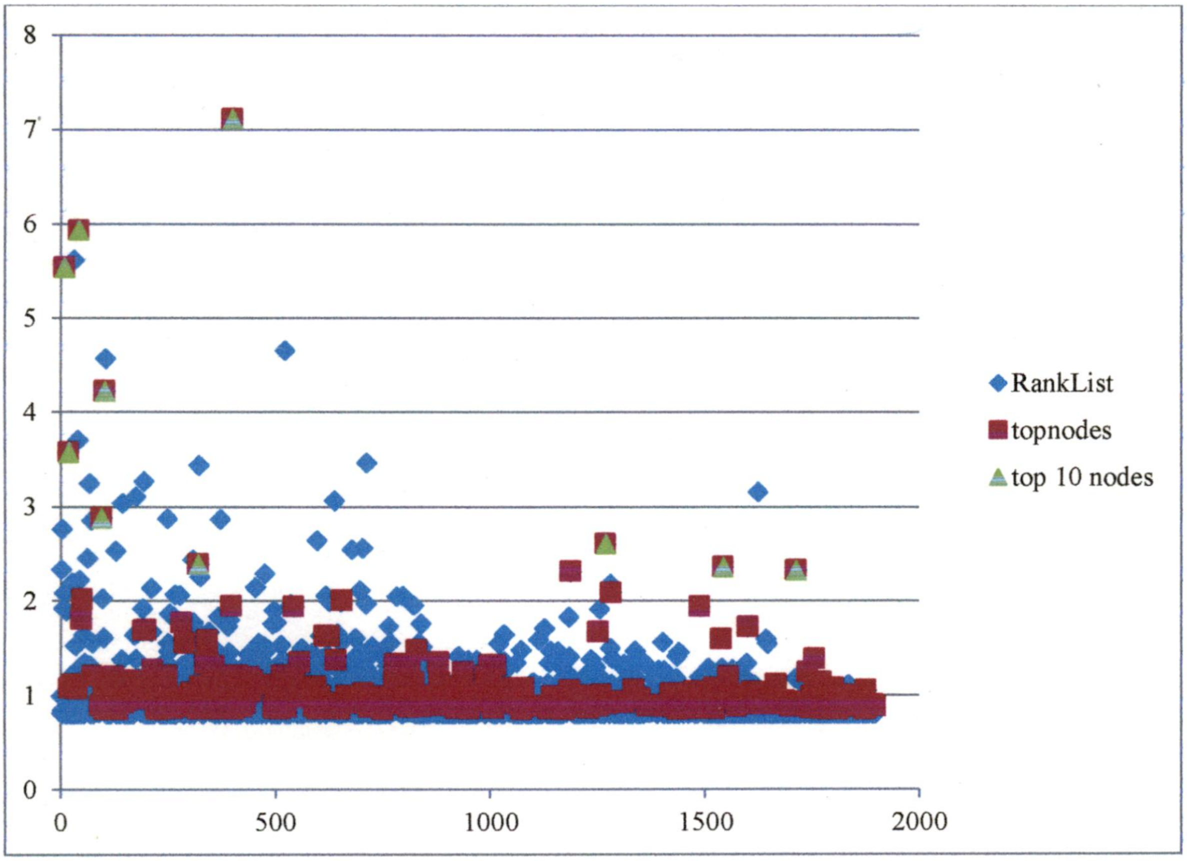

Fig 4.4 Comparison of nodes.

4.3 Results

X-axis --- the nodes number

Y-axis --- node-position value

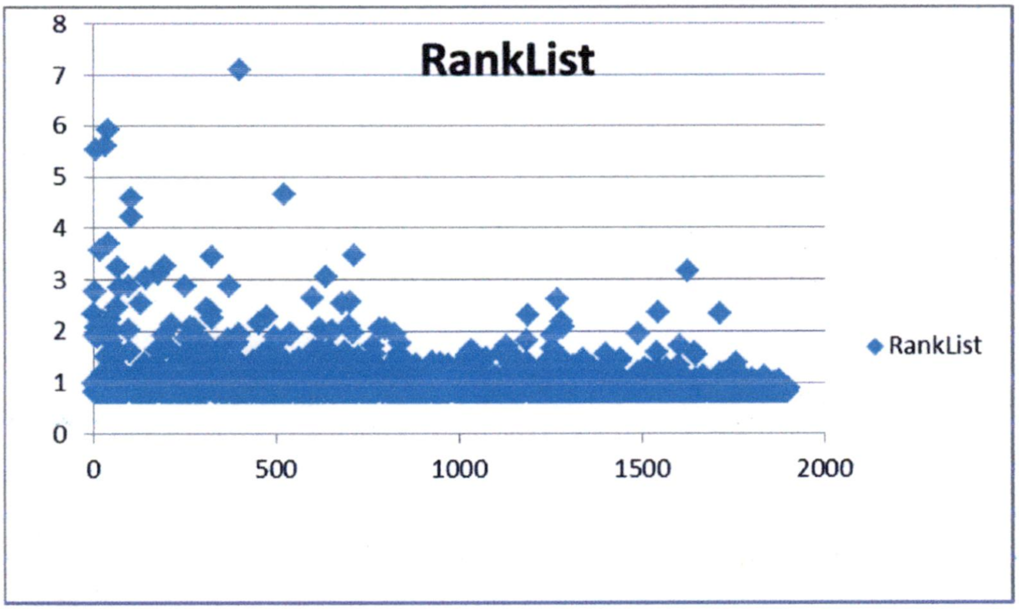

Fig 4.1 *RankList* obtained from node position values

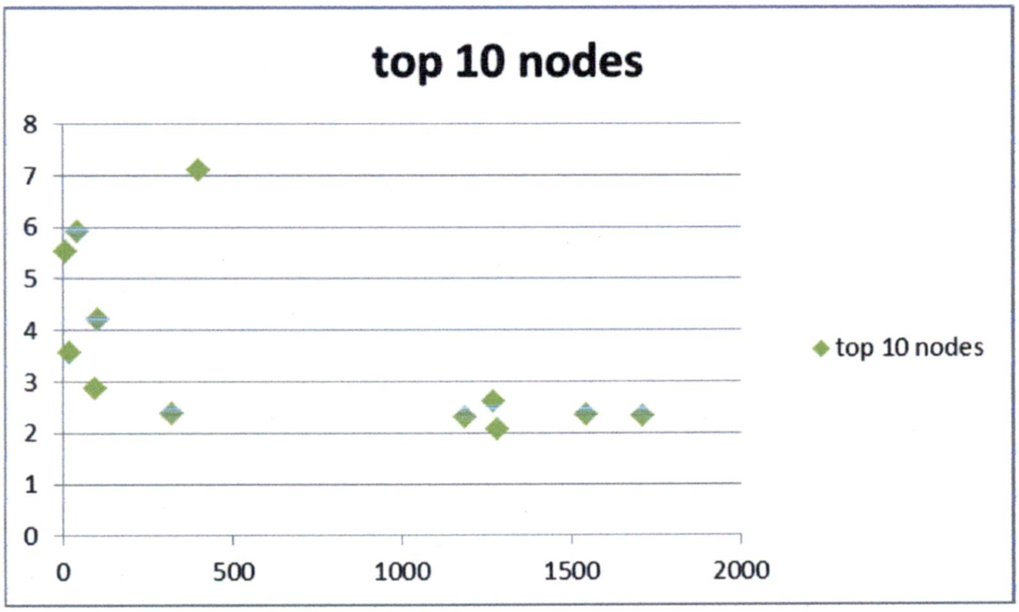

Fig 4.2 Node filtered after applying algorithm with k=10

Time taken for finding top-1000 nodes= 0.031s

Fig 4.4 gives details about which are the node filtered. The blue nodes are the actual ranklist, Red are the node obtained after filtering. Green are the nodes top nodes. as we see these node have high influence value and they are not forming cluster. These nodes are scattered in the network. the nodes are maximize the information dissemination.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

The framework proposed as a part of the dissertation provides an algorithm for computation of influential nodes in social network with considering the time series data as well as ordinary data. It shown that time series data included algorithm is more accurate in computing influential nodes than previous algorithm. The algorithm is compared with other centrality measures. Our algorithm is faster than closeness centrality algorithm. And also influential nodes calculated have high closeness centrality value.

In order to solve the top-k nodes problem we have proposed an algorithm that uses the influential values obtained from the previous algorithm. Algorithm proposed filter the nodes which are not effective in information dissemination.

5.2 Future work:

In real world scenario we need to find the influence of the nodes on fly. So we need faster algorithms for influence calculations. our algorithm is still is not so optimized. Implementing the algorithm in CUDA like parallel processing platforms will be a good contribution to work.

This is work one single relationship like email commination is considered for evaluation of influential nodes. in real world relationships are not single, they are multi in nature. So we need an algorithm which will considers two or more relationship in evaluation of influential nodes.

Last but not least, social network dataset is huge to story in memory, we need a new efficient data structure for storing data special for social network analysis purpose.

REFERENCES

[1] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, San Diego, California, USA session: social network pp 29 – 42, 2007

[2] J. Kleinberg. "*Authoritative Sources in a Hyperlinked Environment*". Journal of the ACM, 46:604–632, 1999.

[3] J. Goldenberg, B. Libai, E Muller,: "Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata". Academy of Marketing Science Review (2001)

[4] K. Musial, P. Brodka and P. Kazienko, "A Performance of Centrality Calculation in Social Networks". In: CASON '09: International Conference on Computational Aspects of Social Networks, 24-27 June 2009. IEEE Press, pp. 24-31, 2009.

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. *Graph Structure in the Web: Experiments and Models*. In Proceedings of the 9th International World Wide Web Conference (WWW'00), Amsterdam, May 2000.

[6] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. Physical Review, 73, 2006.

[7] K. Musiał, P. Kazienko, P. Bródka, "User position measures in social networks", Proceedings of the 3rd Workshop on Social Network Mining and Analysis, p.1-9, June 28-28, 2009, Paris, France.

[8] S. Wasserman, K. Faust. "*Social network analysis: Methods and applications*". New York: Cambridge University Press, 1994

[9] L. Li, D. Alderson, J. C. Doyle, and W. Willinger, "Towards a Theory of Scale-Free Graphs: Definitions, Properties, and Implications". Internet Mathematics, 2(4): pp 431–523, 2006.

[10] R. Hanneman, M. Riddle, "Introduction to social network methods".Online textbook, available from: http://faculty.ucr.edu/~hanneman/nettext/.2006

[11] G. Jennifer, J. Hendler, "Accuracy of Metrics for Inferring Trust and Reputation" in *Proceedings of 14th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2004,LNCS 3257, Springer Verlag,* p116–131,2004

[12] D. Kempe, J. Kleinberg, É. Tardos "Influential nodes in a diffusion model for social networks". The 2005 international colloquium on automata, languages and programming, Springer verlag, pp 1127–1138, 2005.

[13] U. Brandes, T. Erlebach. "*Network Analysis, Methodological Foundations*". Springer – Verlag, Berlin, Heidelberg, Germany,p87-92, 2005.

[14] P. Kazienko." *Associations: Discovery, Analysis and Applications* ."Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej, 2008.

[15] P. Kazienko, K. Musial, A. Zgrzywa. "*Evaluation of Node Position Based on Email Communication*".in Control and Cybernetics. 38 (1), 2009, in press

[16] M. Kimura, K. Yamakawa, K. Saito, H. Motoda, "Community analysis of influential nodes for information diffusion on a social network," *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on* , vol., no.,pp.1358-1363, 1-8 June 2008.

[17] P.A. Estevez, P. Vera, K. Saito, "Selecting the Most Influential Nodes in Social Networks," *Neural Networks, 2007. IJCNN 2007. International Joint Conference on* , vol., no., pp.2397-2402, 12-17 Aug. 2007.

[18] J. Brown, P. Reinegen, "Social ties and word-of-mouth referral behavior". Journal of Consumer Research 14,p350–362,1987.

[19] A. Degenne, M. Forse, "Introducing social networks", London: SAGE Publications Ltd, 1999.

[20] P. Carrington, J. Scott, S. Wasserman,"Models and methods in Social Network Analysis". Cambrige University Press, Cambrige, 2005.

[21] M.E.J. Newma, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality". Physical Review E 64, 016132, 2001.

[22] L.C Freeman, "Centrality in social networks: Conceptual clarification. Social Networks 1", 215-239.1978.

[23] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, "The architecture of complex weighted networks". Proceedings of the National Academy of Sciences 101 (11), 3747{3752}, 2004.

[24] T. Opsahl, P. Panzarasa, "Clustering in weighted networks". Social Networks 31 (2), 155-163. 2009.

[25] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks", in Proc. KDD, 2009, pp.199-208

[26] Y.M. Li; C. Lai; C.L. Hao ; , "Discovering Influential Nodes for Viral Marketing," *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on* , vol., no., pp.1-10, 5-8 Jan. 2009

[27] YFiles Library http://www.yworks.com/en/products.html