MODELING AND ANALYSIS OF SOCIAL NETWORKS

A thesis submitted to the University of Hyderabad in partial fulfillment of the requirements for the award of

Doctor of Philosophy

in

Computer Science

Sreedhar BHhukya

07MCPC17

SCHOOL OF COMPUTER AND INFORMATION SCIENCES

UNIVERSITY OF HYDERABAD

HYDERABAD

(P.O.) CENTRAL UNIVERSITY

HYDERABAD - 500 046, INDIA

December 31, 2013

CERTIFICATE

This is to certify that the thesis entitled "**MODELING AND ANALYSIS OF SOCIAL NETWORKS**" submitted by **Sreedhar Bhukya** bearing Reg. No. 07MCPC17 in partial fulfillment of the requirements for the award of **Doctor of Philosophy in Computer Science** is a bonafide work carried out by him under my supervision and guidance.

The thesis has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Dr. S.Durga Bhavani
Supervisor
School of Computer and Information Sciences
University of Hyderabad

Prof. Arun K.Pujari
Dean
School of Computer and Information Sciences
University of Hyderabad

DECLARATION

I, **Sreedhar Bhukya**, hereby declare that this thesis entitled "**MODELING AND ANALYSIS OF SOCIAL NETWORKS**" submitted by me under the guidance and supervision of **Dr. S.Durga Bhavani** is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Date:

Name: **Sreedhar Bhukya**

Signature of the Student:

Regd. No. **07MCPC17**

ACKNOWLEDGEMENTS

My deepest gratitude to my supervisor **Prof. S. Durga Bhavani** whose valuable guidance and support from preliminary to concluding level enabled me to develop an understanding of the subject. I am thankful to her for her patience and constant encouragement in motivating me with the words of hope throughout this work.

It is my pleasure to thank DRC members, **Prof. S Bapi Raju** and **Dr. Siba Kumar Udgata** for their valuable suggestions in completing my research work.

I would like to express my sincere thanks to Dean, School of Computer and Information Sciences **Prof. Arun K. Pujari** for his cooperation.

I would like to thank **UGC** (University Grant Commission) for providing fellowship for the period of Dt: 23-08-2007 to 22-8-2012 to pursue this research in the field of Social Networks.

I would like to avail this opportunity to thank my parents, **Mr. Bhukya Dharma** and **Mrs. Veeramma** whose good wishes enabled me to pursue and achieve my goal. I gratefully thank my wife **Mrs. Ramya Bhukya** and my two sons **Yashodharan** and **Vasikaran** for their concern and moral support throughout my academics.

I also thank **Prof. H. Mohanty**, **Prof. Arun Agarwal**, **Prof. Chakravarthy Bhagvati**, **Prof. C. R Rao**, **Prof. P. N. Girija**, **Prof.K.N. Murthy**, **Dr. T. Sobha Rani**, **Dr. Anupama P.**, **Mr. Wilson Naik Bhukya** and Staff of School of Computer and Information Sciences, University of Hyderabad.

I extend my special thanks to **Dr. Venkateswar Rao**, Controller oof Examination, **Devesh Nigam**, Deputy Registrar, **Mr. B. Tukaram**, Assistant Registrar, Academic Section, **Mr.B. Srinivas**, Assistant Registrar, Finance section, and employees of the Student Service Section and Finance Section.

I would like thank to **Dr. V. Krishna**, Principal, Govt.Degree and P.G college, Bhadrahalam, **B. Kasya**, **E. Ramakrishana Prasad**, **Eslavath Seetharamulu**, Grand father-in-law, sisters **M. Padma**,**E. Padma**,**B. Kamale**,**B. Rukma**,**B. Mangamma**,**B. Rajakumari**, brother-in-laws **M. Shankar**,**E. Sivaram Prasad**,**B. Mangilal**,**T. Venkanna**, my elder brother **B. Ramulu** and other family members of my sisters, brother and brothers-in-law.

I would like to thank my close friends from School of Computer and Information Sciences **Rusydi Umar**, **Amer Ali Sallam**,**Shakeel Ahmed**, **Dr. S. Mini**, **G. S. Kedarnath**, **Raghu nath Pasunuri** for their moral support.

The warm support of all my friends in University namely, **Akhter Mohiuddin**, **Sanjeev Kumar**, **Aalu Boda**, **Vijay kumar Bhukya**, **L. Vachaya**, **N. Rambabu**, **N. Sreenu**, **Eslavath Bhanu Prasad**, **Eslavath Syam sundar** and other research scholars in our and other departments, who enabled me to complete this thesis and have a wonderful time along the way.

December 2013 Sreedhar Bhukya

ABSTRACT

Networks model physical, biological and social phenomena. In the thesis we focus on network growth models based on preferential attachment and their applicability to real world social networks. Also the important problems of community discovery and influence maximization in social networks are addressed.

Preferential attachment is a process by which a new node joins the existing network by choosing to connect with node(s) of its preference. One of the latest algorithms is that of Toivonen et al. R.Toivonen *et al.* (2006) who proposed a generative model that uses preferential attachment mechanism to form links of the type friend-of-friend (secondary attachment). This model is important since it captures natural friendship formation in the real-world scenario. We observe that in every day life, new friends are made not just through friend-of-friend relationship but also many times via friend-of-friend-of-friend (tertiary attachment). Hence, there is a need to investigate higher order attachment models both mathematically and empirically. We incorporate these ideas and propose a tertiary attachment model.

In this thesis we propose a dynamical network evolution model based on secondary and tertiary attachment mechanisms for generating networks so that they can capture properties of real world networks. A theoretical investigation of the proposed Tertiary Attachment (TA) model is carried out by deriving the equations for rate of change of degree and clustering coefficient for this model. We show that this model follows very closely the model of R.Toivonen *et al.* (2006) in exhibiting important characteristics of social networks like small world property, assortativity, community structure and scale-free property of the network.

We investigate the applicability of the tertiary attachment model thoroughly in the context of two different types of social networks: a) collaboration networks and b) friendship networks. Genetic Programming(GP) data set and Facebook(FB) data set for which time stamps are available are chosen for this study. We show that the

rate equations derived for the TA model, simulated with different parameter distributions, match very closely with the rate of change of degree and clustering coefficient in GP. We also conduct an extensive analysis for triad formation in these networks from the perspective of the predictive capability of the model.

Community discovery is an important challenging problem in social networks. We propose an efficient algorithm for the problem of community discovery by designing an enhancement to the classical algorithm of Girvan and Newman (GN)Girvan and Newman (2002) based on edge-betweeenness score. GN algorithm does not scale very well for large sized networks. In this context, we propose a heuristic that is intended to speed up GN algorithm by reducing the total number of iterations thus avoiding many calculations of the betweenness score. This heuristic algorithm called Enhanced-GN algorithm shows significant improvement over the GN algorithm, along with retaining the accuracy. We show that Enhanced-GN algorithm performs quite well on bench mark data sets and a few real world networks. Additionally, we propose an influence maximization algorithm that utilizes influential nodes within communities for faster influence spread. Influential nodes extracted from the communities discovered by Enhanced-GN algorithm are provided as seeds to the RankedReplace algorithm of Charu C. Aggarwal and Yan (2011). We show that significant improvement in speed is achieved for the RankedReplace algorithm with the proposed heuristic.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

SN	Social network
IC	Initial contact
SC	Secondary contact
TA	Tertiary attachment
m_r	Preferential initial attachment
m_s	Secondary contacts
m_t	Tertiary attachment
CC	Clustering coefficient
PA	Preferential attachment
BA	Barabasi Albert
SF	Scale free networks
FB	Facebook
GP	Genetic programming
Aplen	Average path length
AvgCC	Average clustering coefficient
RR	Rank replacement
RRDD	Rank replacement degree discount
MCL	Markov Cluster Algorithm

NOTATION

Lc	Largest component
n	Number of nodes
m	Number of edges
G	Graph

CHAPTER 1

INTRODUCTION

1.1 Social networks

Social network analysis is a subfield of network science, a study which establishes the structure and function of a network. Social networks are generally modeled as graphs in which the nodes represent individuals or organizations and are connected based on the specific type of relationship they share. Social networks have been intensively studied by social scientists Milgram and Stanley (1967); Granovetter (1973); Wasserman and Faust (1994), for several decades in order to understand local phenomena such as local formation and their dynamics, as well as network wide process, like transmission of information, spreading disease, spreading rumor, sharing ideas etc. Various types of social networks, such as those related to professional collaboration Watts and Strogatz (1998); Newman (1998, 2004*a*), internet dating Holme *et al.* (2004), and opinion formation among people have been studied. Important social network properties include hierarchical community structure Girvan and Newman (2002), small world property Newman (2003), power law distribution of node degrees Krapivsky and Redner (2001).

For a long time, it was believed that networks grew in a random manner. Yule (1925), a mathematician, was the first one who proposed the concept of preferential attachment in the evolution of a network. Preferential attachment is a process in which a new node joins the network by choosing to link with certain existing nodes, exercising a preference rather than joining randomly. Barabasi and Albert (1999) proposed a generative procedure to grow a network, by repeated application of the preferential attachment rule. They showed that a network thus generated exhibits a power law and hence is a scale-free network which is highly observed in nature.

More recently R.Toivonen *et al.* (2006) proposed a generative model using secondary preferential attachment mechanism. Tomassini and Luthi (2007); Luthi *et al.*

(2007) show that networks grown according to secondary attachment rule can be shown to simulate collaboration networks. We extend these ideas and propose a tertiary attachment model.

1.2 Motivation

Most of the models proposed in the literature are based on preferential attachment properties for growing the networks ie. a new node joining the network prefers to link itself to a highly interacting node.

We observe that in every day life, new friends are made not just through friend-of-friend (secondary) relationship but also many times via friend-of-friend-of-friend (tertiary attachment). Hence, there is a need to investigate higher order attachment models both mathematically and empirically.

1.3 Contributions

1.3.1 Novel model for social networks: TA model

In this thesis we propose a dynamical network model based on secondary and tertiary attachment mechanisms for generating networks so that they can capture properties of real world networks. A theoretical investigation of the proposed Tertiary Attachment (TA) model is carried out by deriving the equations for rate of change of degree and clustering coefficient for this model.

1.3.2 Empirical investigation of TA in real world networks

We investigate the applicability of the tertiary attachment model thoroughly in the context of two different types of social networks: a) collaboration networks and b) friendship networks. Genetic Programming(GP) data set and Facebook(FB) data set for which time stamps are available are chosen for this study. We show that the rate

equations derived for the TA model, simulated with different parameter distributions, match very closely with the rate of change of degree and clustering coefficient in GP.

1.3.3 Efficient approach to community discovery

Community discovery is an important challenging problem in social networksYan and Gregory (2012). We propose an efficient algorithm for the problem of community discovery by designing an enhancement to the classical algorithm of Girvan and Newman (GN)Girvan and Newman (2002) for community discovery. GN algorithm involves a measurement called edge betweenness score to be calculated for every edge and the process repeated $|E|$ times, in the worst case, in order to discover the underlying communities in a social network. The computational complexity of GN algorithm is $O(|V||E|)$, hence the execution time is prohibitive for large scale networks. In this context, we propose a heuristic that reduces the total number of iterations thus avoiding many calculations of the betweenness score. This heuristic algorithm called Enhanced-GN algorithm shows significant improvement over the GN algorithm, along with retaining the accuracy. Further, we apply this algorithm to the problem of influence maximization

1.4 Thesis organization

Chapter 1 carries introduction to the thesis.

Chapter 2 gives notation and basic terminology required for social network analysis as well as a general literature study of the different models proposed for social networks.

Chapter 3 describes our proposed tertiary attachment model. The rate equation for growth of average degree of a node as well as average clustering coefficient are

derived mathematically. The model is simulated by varying the parameters and optimal parameters are found which simulate networks that possess small world property, assortativity, power law degree distribution etc.

In **Chapter 4** the tertiary attachment model is investigated for real world bench mark data sets available with time stamp. We consider Genetic Programming(GP) dataset which is a collaboration network and a portion of Facebook data set which is a friendship network. We empirically investigate the presence of tertiary attachment in GP and Facebook. Further, we simulate the tertiary attachment model to grow a network of size of the largest component of GP in order to compute average statistics of the two networks. We find that the average clustering coefficient and average degree statistics are close to those of the GP data set. The analysis using Tertiary attachment model is also carried out for Facebook data set (September-2006 to December-2008).

In **Chapter 5** we propose an improvement to the classical GN algorithm for community discovery. Our proposed algorithm is tested on 6 benchmark data sets and a significant improvement in the running time of the algorithm is achieved. It shown to run more than 50% faster than GN in most cases also preserving accuracy. We investigate further applications of edge betweenness score for the influence maximization problem which is currently a challenging problem.

Chapter 6 ends with conclusions and future directions of the work.

CHAPTER 2

TERMINOLOGY AND BASIC MODELS OF SOCIAL NETWORKS

Building social network graphs is based on parameters whose evolution is statistical in nature. Estimation of these parameters can be achieved in two ways according to traditional sampling theory, namely, design based and model based. In the design based approach elements are selected by a random mechanism to create a sample where as in model based approach, the model specifies the relationship between the sample and the population. In recent decades, both approaches are being used in conjunction with each other. Modeling has two benefits

1. To understand the formation and evolution of social network
2. To simulate and predict the network dependent social process

These can be further subdivided into two classes namely static model and dynamic model. Static models try to explain the properties of a single snapshot of a network where as dynamic models deal with the changes in the network. With the emergence of online networks in these recent years, the study of dynamic models received a great boost.

Static network consists of a fixed number of vertices and grows by only adding a fixed number of edges at every time stamp, a dynamical network starts from a seed and grows at every time stamp with a specified mechanism for attachment until the network grows to a desired size.

2.1 Basic terminology

2.1.1 Graph

A graph $G = (V, E)$ is a visual or symbolic representation of a set of objects V connected by a relation. These objects are called "vertices" or "nodes", each node can be identified by an integer $i = 1, 2, ..., n$. Each link is identified by an unordered pair $\{i, j\}$ that represents a connection between the nodes i and j called "edge" and is graphically represented by a line or curve between the nodes. The links may be labeled or weighted according to the application.

For example, in a collaboration of authors network, nodes represent authors and an edge (i, j) indicates that the authors i and j have a joint publication and a weight on the edge may indicate the number of their common publications or the edge may be labelled by the year of publication. If it is a citation network with nodes as papers and edge represents citation, then the edges are directed making the edge an ordered pair (i, j) indicating the paper i cites paper j.

2.1.2 Degree

Degree of a node i in a graph is the number of nodes connected it and is denoted as k_i. For a directed graph one can define in-degree and out-degree. An isolated node has degree zero.

$$\text{average } deg(G) = \frac{\sum k_i}{n}$$

The degree distributions $P(k)$ of large social networks are often highly skewed, with some nodes having very high degrees.

2.1.3 Adjacency, neighbors

Nodes i and j are said to be adjacent, or neighbors, if $E(G)$ contains an edge i, j. $E(G)$ without multiple links can be represented as an adjacency matrix A, in which the elements $a_{ij} = a_{ji} = 1$ if $i, j \in E(G)$, and $a_{ij} = a_{ji} = 0$ otherwise.

2.1.4 Component

A component of a network is a maximal connected subgraph, we often consider the largest component of a graph. The size of the largest component is denoted by Lc (Largest component).

2.1.5 Clustering coefficient

Clustering coefficient is the measure of the probability that the adjacent nodes of a given node are connected. Average clustering coefficient is the ratio of number of edges between the neighbors of a given node to the maximum number of edges possible. For an undirected graph this is given by,

$$CC(i) = \frac{2E_i(k_i)}{k_i(k_i-1)}$$

Where $E_i(k_i)$ is the number of triangles at the given node i and the factor $k_i(k_i - 1)/2$ is the maximum number of triangles incident at i in an undirected graph. For a directed graph it will be $k_i(k_i - 1)$.

$$Avg(CC) = \frac{\sum_{i=1}^{n} CC(i)}{n} \tag{2.1}$$

Note that $Avg(CC) = 1$ for a complete graph and is 0 if the graph is completely disconnected.

2.1.6 Average path length

Average path length is the average number of edges presents in the shortest path between any two nodes of a network. The shortest path between any two nodes is called $geodesic$. All geodesics between a given pair of nodes will have same length, by definition. If d_{ij} is the length of the shortest path between the nodes i and j of a network, then the average path length $Aplen$ of an undirected graph of n nodes is

given by,

$$Aplen = \frac{1}{n(n-1)} \sum_{i \geq j} d_{ij}$$

2.1.7 Scale-free network

This is a network whose degree distribution follows power law. $P(k) \sim k^{-\gamma}$ with exponent $2 < \gamma < \infty$.

For example one can say average human life-time is 60–80 years. That is, the scale of average human life-time is of the order of a few decades. But for long tailed distributions which obey power law distribution, one can not attribute a scale and hence these are called scale-free networks. Many real world networks including biological networks, world-wide web links and social networks are scale-free networks.

2.1.8 Preferential attachment

This is the process in which a new node entering the graph prefers to attach to a node with higher degree. Even though the selection is random, there will be a bias towards the node with higher degree during selection. Thus the probability that a new node connects to a given node is given by $k(i)/\sum_j k_j$.

Preferential attachment was first proposed by Yule (1925) in his "A Mathematical Theory of Evolution" which is called Yule process. A process using this rule leads to a scale free network proposed by Barabasi and Albert (1999) jn "Emergence of scaling in random networks".

2.1.9 Vertex correlation and assortativity

In the preferential attachment model, a new node prefers to attach to a node with higher degree. In this case there is a tendency for higher degree vertices to connect with other high degree vertices. This phenomenon is called assortative mixing or assortativity.

2.1.10 Small-world network

Small world network is a type of network in which the average path length is of the order of logarithm of number of nodes in the network. Small world networks have high clustering coefficient and small average path length.

Most of the real world networks like food chains, electric power grid, social networks, random networks are small world networks Newman (2000).This concept was introduced by Milgram and Stanley (1967) after their investigating "Six degrees of separation".

2.1.11 Diameter of graph

The longest of the shortest paths among all the pairs of nodes in a graph is called the diameter of the graph. ie; it is the maximum distance among all pairs of vertices.

In the real world scale-free networks, the diameter is expected to be small and the clustering coefficient very large.

2.1.12 Community structure

A graph is said to have community structure if the nodes can be separated into groups which are tightly knit within but with sparse inter-group connections.

Discovering community structure enables us to understand the topology of the network and distribution of information among various nodes.

2.2 Basic models for social networks

Network models can be classified as static models and dynamic models. Static models have fixed number of nodes and a fixed number of edges chosen by a specified rule. The dynamic models typically have a growing network adding nodes and edges at every time step.

The Erdos-Renyi-Gilbert Erdos and Renyi (1960) model is a static random graph model for an undirected graph, given a fixed number of nodes and a fixed number of edges (E) chosen randomly. The "Random graphs" model was proposed by Gilbert (1959) independently at the same time and hence the name to the model. They have studied the properties of the model as 'E' increases, and further extensions of this random graph model were achieved by Airoldi (2006); Maria (2005); M (2007); BollobÃąs *et al.* (2007).

In the social science community, the developments started with the invention of sociogram by Moreno (1934). Mathematical formulation to this model was given in terms of matrices and graphs and the analysis carried out by Luce and Perry (1949); Luce (1950); Luce *et al.* (1955); Radner and Tritter (1954); Spilerman (1966); Alba (1973).

Famous extensions of Erdos-Renyi -Gilbert model with a fixed number of edges have been proposed by rewiring the edges in two ways: one is the preferential attachment model Barabasi and Albert (1999) and another is the small world model Watts and Strogatz (1998). These two models were led to a surge of interest in network science due to their ability to imitate many real world networks Barabasi and Albert (1999); Newman (2004*b*); Chung and Lu (2006). Further extensions have been done by the statistical physics models which lead to Monte Carlo Markov Chain (MCMC) models Blitzstein and Diaconis (2006); Handcock *et al.* (2008).

Newman *et al.* (2002) started with a model based on random graphs of Erdos-Renyi and attached degrees to each node randomly in accordance with the real network data to produce a random graph whose degree distribution turned out to be similar to that of real world networks. They studied the properties of clustering coefficient, average degree and average path length for these random graphs and compared with the real network data. The results are found to compare favourably. Thus this model paved a way for the usage of random graphs to real world social networks.

Kanovsky (2010) proposed Extended Watts and Strogatz (EWS) model, which is the simplest model for small-world of social networks and studied some of the properties like small world phenomenon and clustering coefficient. This is one of the

simplest models that incorporates the basic properties of a small world network and they proposed an algorithm for community recognition in social networks.

One of the first probability models was initiated as p1 model proposed by Holland and Leinhardt which is a generalization of Erdos-Renyi-Gilbert model for directed graphs. Holland and Leinhardt (1981). To address the issue of large number of parameters in the p1 model, p2 model was developed Wasserman and Pattison (1996). Later Frank and Strauss (1986) started with an assumption that two edges are dependent only if they have a common node and proposed a Markov graph model and is further generalized by Wasserman and Pattison (1996) with the name "Exponential random graphs model" or p* model. Snijders et al. pioneered work on building classes of probability models and also carried out evaluation of various stochastic modelsSnijders (2005, 2006). Wasserman and Pattison (1996); Handcock (2003); Goodreau (2007) studied Monte Carlo MC methods in simulating large networks and showed up the problem of instability in these networks.

It is interesting to focus on network models that choose transitivity as a main property for evolution. Granovetter proposed a mechanism of triadic closure, a rule by which two nodes tend to form a link due to the presence of a common neighbour. One of the first stochastic models is a Markov graph model due to Frank and Strauss, 1986 who maintain the the triangle count while evolving the network. Kumpula *et al.* (2007) proposed a model in which the evolution mechanism involves both triadic closure of links and global connections between randomly picked nodes. R.Toivonen *et al.* (2006) explicitly include a stochastic rule that creates triangular connections, the nodes involved are referred to as secondary contacts.

Algorithmic strategies Arabie *et al.* (1978); Doreian *et al.* (2004); Fienberg (2007) and statistical physics and computer models Kernighan and Lin (1970); Clauset (2005); Newman (2006); Shalizi *et al.* (2007); Mishra *et al.* (2008) have been proposed. Now a days the availability of data for dynamical analysis, and advances in statistical and computational networks lead to the development of dynamical model social networks.

Some of the other important models in dynamical analysis are continuous markov

model(CMPM)Holland and Leinhardt (1977); Wasserman (1977), discrete time markov model of Banks and Carley (1996); J.M. (2002) and discrete markov ERGM model of Hanneke and Xing (2007) which are well discussed in literature.

We propose a model extending the secondary attachment model of Toivonen et al. proposing a stochastic rule based on what we call tertiary contacts which is described in detail in the next chapter.

CHAPTER 3

PROPOSED TERTIARY ATTACHMENT MODEL FOR SOCIAL NETWORKS

3.1 Introduction

Different types of network models have been proposed to study the growth process, diffusion and retrieval of information. A number of recent studies on social networks are based on characteristics which include assortative mixing, high clustering, short average path length, broad degree distributions and the existence of community structure.

3.2 Literature on random and preferential attachment

One of the most important models of growing networks is the Barabasi-Albert (BA) model, in which a new node j connects to an existing node i in the network using a preferential attachment rule.

Scientists believed that preferential attachment is a natural process by which growth is achieved in a social network, but to establish this phenomenon one needs data with time stamp. Newman (2001) built scientific collaboration network based on research interaction at Santa Fe institute and included year of publication. They empirically established that the probability of scientific collaboration increases with the number of common collaborations. This led to a large body of work in social network with models proposed based on preferential attachment (Jeong *et al.*, 2003; Vázquez, 2003; Flaxman *et al.*, 2004; Capocci *et al.*, 2006; Wang *et al.*, 2008)

Catanzaro *et al.* (2004) proposed assortative model to study the scientific collaboration in arxiv.org. The model is based on preferential attachment with assortative

degree distribution. They have studied the properties as a function of a parameter p, the probability of adding a new node to an existing node. Thus this model studied the microscopic behavior of the social network based on preferential attachment model.

Fenner *et al.* (2007) proposed a stochastic model to study the interactions in social networks where users may join, deactivate and reactivate themselves according to preferential attachment. They have proposed a meanfield model which gives power law degree distribution asymptotically. This model is well applicable to users of WLAN networks and peer-to-peer networks.

Von Arb *et al.* (2008) proposed a Friend-of-Friend detection in mobile social networking scenario. Two persons compare their address books without revealing their own contacts except for those that match. In other words they want to find their common friends without revealing their individual list of contacts. If they find a common friend they may interact with each other through this friend-of-friend relationship. This system may be quite useful when one goes to a new place.

Wei Deng *et al.* (2012) proposed a mechanism of growth in which preference algorithm is based on node weights in online social networks. They study the evolution model and conduct simulation on online social networks where they analyze how user connection weights increase over period of time.

Buscarino *et al.* (2012) proposed another attachment mechanism, based on the idea that a new node links to the nodes of a community, in addition to randomly picked nodes. Their idea is that friends and communities should be treated on different grounds. One cannot ensure that all friends have same interests. So community selection should be considered separately independent of friends selection. This model is well applicable to the real world networks like Facebook etc.

Ball and Newman (2012) considered a directed network among the US high school and junior high school students. They have observed that the preferences depended on social status, with unreciprocated friendships being far from random. It was found that the lower ranked individuals seek friendships of higher ranked individuals and that decides the direction of interaction (edge) in the graph.

Toivonen *et al.* (2009) described a classification study of network growth models into mainly two types, namely, Network Evolution Model (NEM) where the addition of new edge depends on the local structure of the network and the Nodal Attribute Model (NAM) where new links are generated based on the properties of the node. It was observed that NEM gave good degree distribution and clustering coefficient and NAM gave good assortativity and community structure when compared with the real world networks.

R.Toivonen *et al.* (2006) proposed secondary attachment model as an improvement to BA model. Their model is stated quite simply by incorporating random attachment as well as preferential attachment based on friend-of-friend preferential rule. Our present work is an extension of these preferential attachment models. We have extended their mechanism with an additional tertiary attachment for a better description of the network.

We propose a tertiary attachment model for social network which satisfies all the important characteristics like. small world property, assortative mixing, high clustering coefficient and low average path length. The growth of the networks is performed using a mixture of random attachment and implicit preferential attachment.

3.3 Motivation to tertiary attachment

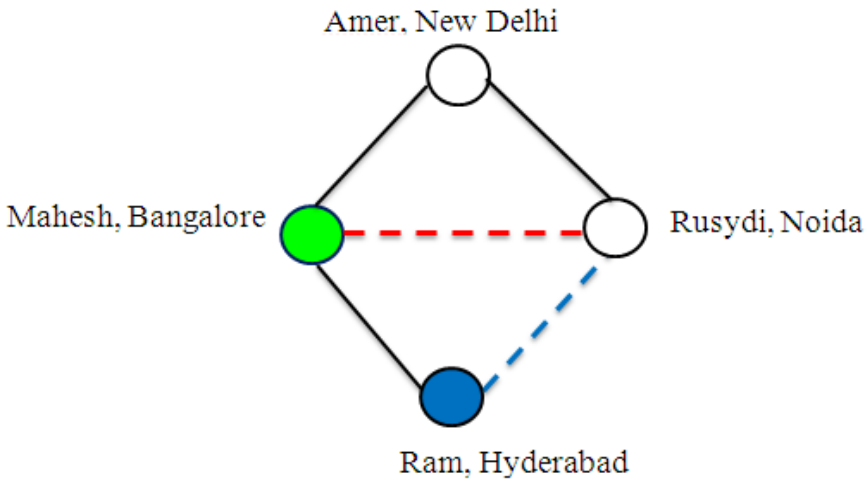

Figure 3.1: Basic idea of tertiary attachment model.

Let us say Ram from Hyderabad contacts his friend Mahesh, his **initial** contact at Bangalore to find information about Noida. Mahesh does not know about Noida city, he contacts his friend Amer at New Delhi to find information of Noida. Amer introduces his friend Rusydi to Mahesh who in turn gives this contact to Ram. That is Mahesh forms a 'friend-of-friend' connection with Rusydi which can be termed as a **secondary** contact. In this process Ram forms a 'friend-of-friend-of-friend' connection to Rusydi which we refer to as **tertiary** attachment.
Note: There is an internal connection that forms between Mahesh and Rusydi. Further, it is possible that secondary attachment forms between Ram and Amer as friend-of-friend.

3.3.1 Example from Indian context

Let us consider a social problem in the context of Indian arranged marriages, where a family head A contacts a friend B for searching bride or bridegroom. Suppose that friend B or B's neighbour did not get the information required by A, but B may get required bride or bridegroom information from neighbour of neighbour families. If B gets the information from neighbour of neighbour familes, then A proposes to form a connection to B's neighbour of neighbour of families for forming relationship, which we consider as tertiary contact. This scenario is quite a common occurence in Indian families searching for bride or bridegrooms. If a person contacts us for some purpose and we are unable to help her, we will try to help by some contacts of our friends.

Model of Toivonen et al. can be naturally extended to incorporate these ideas and we propose a model that reflects this process and name it tertiary attachment model. In the model of R.Toivonen *et al.* (2006), information about friends only needs to be updated, where as in our model, information about friend of friend will be updated. Of course this model creates a complex social network but, sharing of information or data will be very fast. This fulfills the actual purpose of social networking in an efficient way with a faster growth rate by keeping the community structure intact.

3.4 Basic terminology

We adopt the approach of network evolution model (NEM) in which the network structure evolves according to a defined set of stochastic rules. We assume that at each time step, one new node joins the network. In addition to the basic network terminology described in chapter 1, here we are going to introduce terms that are exclusively related to our model.

3.4.1 Initial contact

The vertices present in the network to which a new node joining the network chooses to form connection with, are called initial contacts. These nodes are in general chosen randomly. Thus the probability of selecting an initial contact at any time will depend on the number of nodes already existing in the network at that time.

3.4.2 Secondary contact

Secondary contact (SC) is a neighbour of initial contact with which the newly entering node makes a contact. This word was introduced by R.Toivonen *et al.* (2006). Barabasi and Albert (1999) treat such connection as a friend-of-friend preferential attachment.

3.4.3 Tertiary contact

We introduce the term tertiary contact (TC) to mean a neighbour of secondary contact or a neighbour of neighbour of initial contact with which the newly entering node makes a contact at a later stage. This is a new concept which we have introduced as an extension to the concept proposed by R.Toivonen *et al.* (2006). This too comes under preferential attachment.

3.4.4 Terminology for edges

Initial attachment is an edge that connects the newly joined vertex to an initial contact. **Secondary attachment** is an edge connecting the new vertex to a randomly selected neighbour of initial contact in the existing network. **Tertiary attachment** is an edge connecting the new node to a tertiary contact.

We now describe a preferential attachment model due to R.Toivonen *et al.* (2006) which we extend by incorporating tertiary attachments. The rate equations for degree distribution and clustering coefficient for this new model are derived and simulations are carried out for different paramters. Average statistics of the networks thus obtained are compared with the model of R.Toivonen *et al.* (2006).

3.5 Secondary attachment model

In this section, the preferential attachment model of R.Toivonen *et al.* (2006) is reviewed.

The algorithm consists of two growth processes, namely, (a) Random attachment and (b) Implicit preferential attachment.

1. Start with a seed network of N vertices.
2. The number of vertices that are picked randomly as initial contacts is $m_r \geq 1$ on average.
3. Pick on average $m_s \geq 0$ neighbors of each initial contact as secondary contacts.
4. Connect the new vertex to the initial and secondary contacts.
5. Repeat the steps 2-4 until the network has grown to desired size.

On average the number of edges possible at a new user $= m_r + m_r m_s$ which includes initial and secondary attachments. Therefore, if 't' nodes have joined the network then, on average, the increase in the total degree of vertices in the network is, $d = \sum k_i = 2m_r(1+m_s)t$, k_i is the degree of vertex v_i. Toivonen et al. propose a rate equation for the degree of a vertex v_i denoted as k_i, as follows:

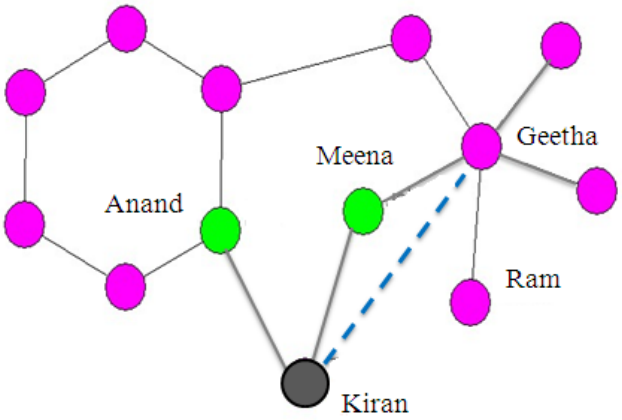

Figure 3.2: $Kiran$ is a new user, chooses two random initial contacts ($m_r = 2$) as $Anand$ and $Meena$. $Geetha$, a neighbour of initial contact is chosen randomly as a secondary contact. On ($m_s = 1$) thick lines from $Kiran$ show initial attachment and the dotted line indicates secondary attachment.

$$\begin{aligned}\frac{\partial k_i}{\partial t} &= m_r\left(\frac{1}{t}+m_s\frac{k_i}{\sum k_i}\right)\\ &= \frac{m_r}{t}+\frac{m_r m_s}{2m_r(1+m_s)t}k_i\\ &= \frac{1}{t}\left(m_r+\frac{m_s}{2(1+m_s)}k_i\right)\end{aligned}$$

Further, Toivonen et al. derive the degree distribution and clustering coefficient to obtain the following equations R.Toivonen *et al.* (2006).

$$k_i(t) = B\left(\frac{t}{t_i}\right)^{\frac{1}{A}} - C \tag{3.1}$$

where $A = 2(1+m_s)/m_s$, $B = m_r(A+1+m_s)$ and $C = Am_r$. The probability density distribution for degree k is given by

$$P(k) = AB^A(k+C)^{\frac{-2}{m_s}-3} \tag{3.2}$$

where the distribution obeys the power law $p(k) \sim k^{-\gamma}$ with $\gamma = 3+2/m_s, m_s > 0$, with exponent $3 < \gamma < \infty$. Let $E_i(k_i)$ denote the number of triangles at v_i of degree k_i. The clustering coefficient is given by

$$C_i(k_i) = \frac{2E_i(k_i)}{k_i(k_i-1)} = \frac{2k_i + D\log(k_i+C) - F}{k_i(k_i-1)} \tag{3.3}$$

where $C = Am$, $D = C(m_s - 1)$ and $F = D\log B + m_r$. Note that clustering coefficient depends on k as $c(k) \sim 1/k$.

We extend the secondary attachment model to include tertiary attachment, which we claim, reflects the reality more closely. We propose tertiary attachment algorithm and subsequently, along the lines of Toivonen et al., we derive the rate equation for the degree of vertex, probability density distribution and clustering coefficient for the proposed model.

3.6 Proposed algorithm

The new model includes three processes: (1) Random attachment (2) Implicit preferential connection with the neighbors of initial contact (3) Tertiary attachment a connection between the initial contact to its neighbor of neighbor is contact.

Algorithm 1 Let $m_r \geq 1$, $m_s \geq 0$ & $m_t \geq 0$ be three probability distributions which are given as input parameter to the algorithm along with the input graph.

procedure (Preferential attachment)
  Initialize a network $G = (V, E)$ with one node
  while $(size \neq Size)$ **do**
    Choose a new vertex t to join the network
    $V = V \cup \{ t \}$
    Choose randomly on average $m_r \geq 1$ vertices in V as initial contacts
    for each initial contact $u \in V$ **do**
      $e(t,u) = 1$ ▷ Edge connects to initial contact
      Choose on average $m_s \geq 0$ vertices v adjacent to u as sec. contact
      for each secondary contact $v \in V$ **do**
        $e(t,v) = 1$ ▷ Edge connects to secondary contact
        Choose on average $m_t \geq 0$ vertices w adjacent to v as ter. contact
        for each tertiary contact $w \in V$ **do**
          $e(u,w) = 1$ ▷ u forms friend-of-friend connection with w
          $e(t,w) = 1$ ▷ Edge connects to tertiary contact
        end for
      end for
    end for
    $size = size + 1$
  end while
end procedure

The growth of the network according to this algorithm is depicted in figure: 3.3.

It can be seen in figure: 3.4 that it is possible in a relatively short time for two initial contacts to establish interaction, whereas in the secondary attachment model, it takes much longer time.

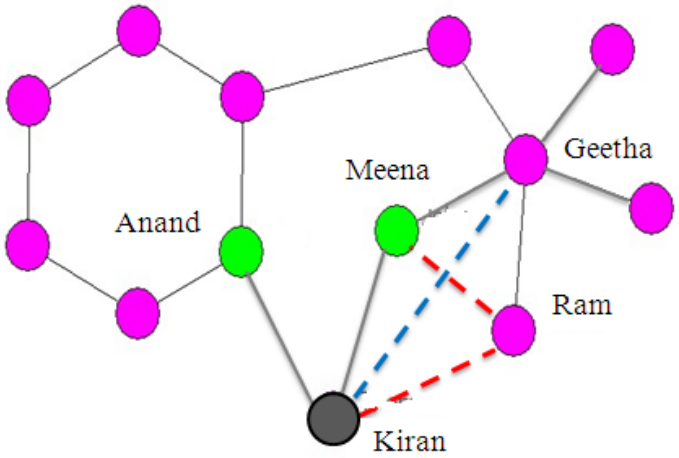

Figure 3.3: The new user $Kiran$ initially connects to initial contacts which are $Anand$ and $Meena$ (green colored vertices). Now user $Meena$ updates its neighbor of neighbor contacts list and hence connects to Ram. Thus $Kiran$ makes secondary and tertiary connections with $Geetha$ and Ram respectively.

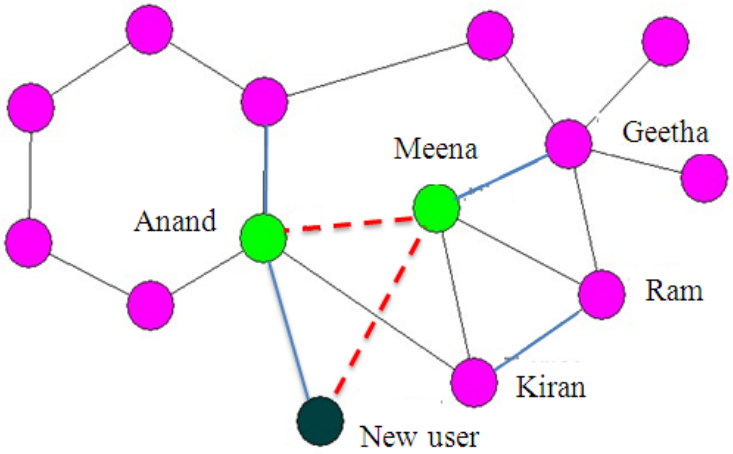

Figure 3.4: Interaction between two initial contacts: Suppose a new user (Block colored vertex) selects either $Anand$ or $Meena$ as its preferential initial contacts (green colored) which are the members of two exclusive communities. There establishes a contact between $Anand$ and $Meena$ via $Kiran$ as well as the new user connects to user $Meena$; both contacts formed as tertiary connections (red colored edges). In the earlier model there is a very small possibility to establish a connection between $Anand$ and $Meena$ at any point of time.

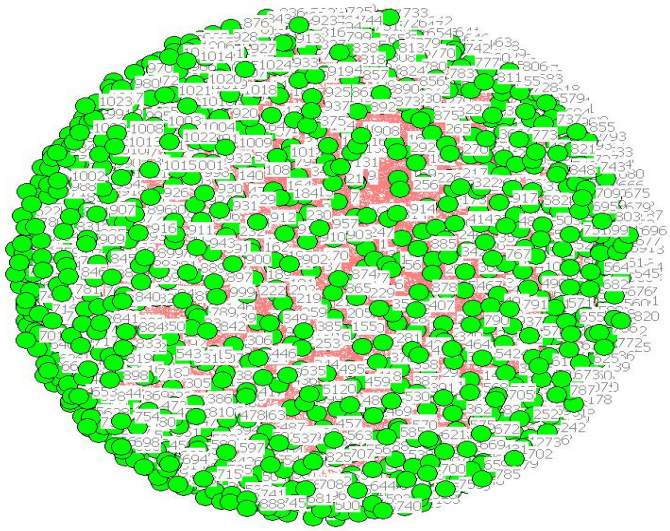

Figure 3.5: Network showing growth upto 1000 vertices with tertiary attachment model simulated in [UCINET 6].

In figure:3.5, a snapshot of the network of size 1000 nodes grown using tertiary attachment model is given.

In the next section, we derive equations for the rate of change of degree of a node as well as rate of change of clustering coefficient at a node on similar lines given by Tovionen et al.

3.7 Rate equation for degree distribution

In order to frame the rate equation that describes how the degree of a vertex changes on average during one time step, we need to look at the processes that contribute to the network growth. We abuse the notation a bit by taking t for time as well as when used as suffix, to stand for tertiary connection.

1. At every time step, one node joins the network. When a new vertex directly links to v_i at any time t, there will be on average t vertices in the network. Here we are selecting m_r out of them as initial contacts with a probability m_r/t.
2. When a vertex links to v_i as secondary contact, the selection will give rise to secondary preferential attachment. These will be $m_r.\ m_s$ in number.
3. When a vertex links to v_i as tertiary contact, this will also be a random preferential attachment. These will be $2m_r m_s m_t$ in number, since the internal edge also is connected as part of tertiary attachment.

$$k_{init} = m_r + m_r m_s + 2m_r m_s m_t$$

Total degree for t vertices is

$$d = 2(m_r + m_r m_s + 2m_r m_s m_t)t$$

The rate equation for rate of change of average degree of a vertex of R.Toivonen *et al.* (2006) is logically extended to include tertiary contacts as given below.

$$\frac{\partial k_i}{\partial t} = \frac{1}{t}\left(m_r + \frac{m_r m_s + 2m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)} k_i\right) \tag{3.4}$$

Let

$$A = \left(\frac{2(m_r + m_r m_s + 2m_r m_s m_t)}{(m_r m_s + 2m_r m_s m_t)}\right)$$

Then

$$\frac{\partial k_i}{\partial t} = \frac{1}{t}\left(m_r + \frac{k_i}{A}\right)$$

Separating and integrating (from t_i to t, and from k_{init} to k_i) we will get the following time evolution for the vertex degrees

$$\begin{aligned} \int_{t_i}^{t} \frac{1}{s}\partial s &= \int_{k_{init}}^{k_i} \frac{1}{(m_r + \frac{k_i}{A})}\partial k_i \\ &= A\log\left(m_r + \frac{k_i}{A}\right)|_{k_{init}}^{k_i} \\ &= A\log\left(\frac{m_r + \frac{1}{A}k_i}{m_r + \frac{1}{A}k_{init}}\right) \\ i.e\left(\frac{t}{t_i}\right)^{\frac{1}{A}} &= \frac{m_r + \frac{1}{A}k_i}{m_r + \frac{k_{init}}{A}} \end{aligned} \tag{3.5}$$

Since

$$\begin{aligned} \frac{k_{init}}{A} &= \left(\frac{(m_r + m_r m_s + 2m_r m_s m_t)(m_r m_s + 2m_r m_s m_t)}{2(m_r + m_r m_s + 2m_r m_s m_t)}\right) \\ &= \frac{m_r m_s + 2m_r m_s m_t}{2} \end{aligned}$$

Substituting in equation 3.5

$$\begin{aligned} k_i &= A\left\{\left(m_r + \tfrac{k_{init}}{A}\right)\left(\tfrac{t}{t_i}\right)^{\frac{1}{A}} - m_r\right\} \\ k_i &= A\left(\tfrac{2m_r + m_r m_s + 2m_r m_s m_t}{2}\right)\left(\tfrac{t}{t_i}\right)^{\frac{1}{A}} - Am_r \end{aligned}$$

$$\begin{aligned} \text{Let} \quad B &= A\left(\tfrac{2m_r + m_r m_s + 2m_r m_s m_t}{2}\right) \\ \text{and} \quad C &= Am_r \end{aligned}$$

Then finally the equation becomes

$$k_i(t) = B\left(\frac{t}{t_i}\right)^{\frac{1}{A}} - C \tag{3.6}$$

From time evolution of vertex $k_i(t)$, we can calculate the degree distribution $p(k)$ by forming cumulative distribution $F(k)$ and differentiating with respect to k. The fraction of vertices whose degree is less than $k_i(t)$ at time t is equivalent to the fraction of vertices that are introduced after time t, since t is evenly distributed, this fraction is $(t - t_i)/2$. These facts lead to the cumulative distribution F as

$$F(k_i) = P(k \leqslant k_i) = P(t \geqslant t_i) = \frac{1}{t}(t - t_i) \tag{3.7}$$

Then

$$\begin{aligned} \left[\tfrac{k_i + C}{B}\right]^A &= \left(\tfrac{t}{t_i}\right) \\ t_i &= B^A (k_i + C)^{-A} t \end{aligned}$$

from 3.5 and inserting it into 3.7, differentiating $F(k_i)$ with respect to k_i, and replacing the notation k_i by k in the equation, we get the probability density distribution for the degree k as

$$\frac{\partial F}{\partial k_i} = AB^A (k_i + C)^{-A-1}$$

$$P(k) = AB^A (k + C)^{\frac{-2}{m_s + 2m_s m_t} - 3} \tag{3.8}$$

Here A, B and C are as above. In the limit of large k, the distribution becomes a power law $P(k) \sim k^{-\gamma}$ with $\gamma = \left(3 + \frac{2}{m_s+2m_s m_t}\right)$ leading to $3 < \gamma < \infty$, Hence the lower bound to the degree exponent is 3. Although the lower bound for degree exponent is same as earlier model, the exponent for the earlier model is $\gamma = 3 + 2/m_s$.

3.8 Rate equation for clustering coefficient

The clustering coefficient on vertex degree can also be found by the rate equation method of Szabo *et al.* (2003). Let us examine how the number of triangles E_i change with time. We follow the derivation of R.Toivonen *et al.* (2006) very closely, only extending the logic to include tertiary attachments. The triangles E_i around v_i are mainly generated in three ways.

1. Vertex v_i is chosen as one of the initial contacts with probability m_r/t and new vertex links to some of its neighbors as secondary contact, giving rise to a triangle.
2. The vertex v_i is chosen as secondary contact and the new vertex links to it giving rise to a triangle.
3. The vertex v_i is chosen as tertiary contact and the new vertex links to it as its tertiary and initial contact links to it as secondary contact and due to an internal update, giving rise to 3 triangles.

These three processes are described by the rate equation as follows:

$$\frac{\partial E_i}{\partial t} = \frac{\partial k_i}{\partial t} - \left(\frac{m_r}{t} + \frac{m_r m_s}{t} + \frac{3m_r m_s m_t}{t} + \frac{5m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)\, t} k_i \right) \quad (3.9)$$

$$= \frac{\partial k_i}{\partial t} - \frac{1}{t} \left(m_r + m_r m_s + 3m_r m_s m_t + \frac{5m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)} k_i \right) \quad (3.10)$$

Let

$$\begin{aligned} a &= -(m_r m_s + 3m_r m_s m_t + m_r) \\ b &= -\frac{5m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)} \end{aligned}$$

Then

$$\begin{aligned}\frac{\partial E_i}{\partial t} &= \frac{\partial k_i}{\partial t} + \frac{a+bk_i}{t} \\ \frac{(\partial E_i - \partial k_i)}{\partial t} &= \frac{a+bk_i}{t}\end{aligned}$$

Integrating with respect to t we get

$$\begin{aligned}\int \tfrac{1}{t}\partial t &= \tfrac{1}{a+bk_i}\int_{E_{init}}^{E_i} \partial E_i - \int \tfrac{\partial k_i}{a+bk_i} \\ \log\left(\tfrac{t}{t_i}\right) &= \tfrac{1}{a+bk_i}(E_i - E_{init}) - \tfrac{1}{b}\int_{k_{init}}^{k_i} \tfrac{b}{a+bk_i}\partial k_i \\ \text{since, } \tfrac{-1}{b}\int_{k_{init}}^{k_i} \tfrac{b}{a+bk_i}\partial k_i &= -\tfrac{1}{b}\log(a+bk_i)|_{k_{init}}^{k_i} \\ &= -\tfrac{1}{b}\log\left(\tfrac{a+bk_i}{a+bk_{init}}\right) \\ \log\left(\tfrac{t}{t_i}\right) &= \tfrac{1}{a+bk_i}(E_i - E_{init}) - \tfrac{1}{b}\log\left(\tfrac{a+bk_i}{a+bk_{init}}\right) \\ \tfrac{1}{a+bk_i}[E_i - E_{init}] &= \log\left(\tfrac{t}{t_i}\right) + \tfrac{1}{b}\log\left(\tfrac{a+bk_i}{a+bk_{init}}\right)\end{aligned}$$

$$E_i(t) = (a+bk_i)\log\left(\frac{t}{t_i}\right) + \left(\frac{a+bk_i}{b}\right)\log\left(\frac{a+bk_i}{a+bk_{init}}\right) + E_{init} \tag{3.11}$$

Now making use of the equation 3.6 to solve for $\log(t/t_i)$ in terms of k_i and inserting it into 3.11 to get $E_i(k_i)$

$$E_i(k_i) = (a+bk_i)A\left[\log(k_i+C) - \log B\right] + \left(\frac{a+bk_i}{b}\right)\log\left(\frac{a+bk_i}{a+bk_{init}}\right) + E_{init} \tag{3.12}$$

where

$$\begin{aligned}E_{init} &= m_r m_s(1+3m_t) \\ A &= \left(\tfrac{m_r+m_r m_s+2m_r m_s m_t}{m_r m_s+2m_r m_s m_t}\right) \\ B &= A\left(m_r + \tfrac{m_r m_s+2m_r m_s m_t}{2}\right) \\ C &= Am_r\end{aligned}$$

Dividing $E_i(k_i)$ by the maximum possible number of triangles, $k_i(k_i-1)/2$, we arrive at the average clustering coefficient.

$$c_i(k_i) = \frac{2E_i(k_i)}{k_i(k_i-1)} \tag{3.13}$$

Since $E(k) \sim f(k\log k)$, for large values of degree k, the clustering coefficient

$c(k) \sim \log k/k$. The value of clustering coefficient is larger compared to the earlier model where $c(k) \sim 1/k$.

The tertiary attachment model is simulated with fixed parameter values in order to analyze some of the network characteristics like degree distribution, clustering coefficient, average path length etc. which are presented in the next section.

3.9 Numerical simulation of tertiary attachment model

The parameters of choice for the tertiary attachment model are probability distribution for initial, secondary and tertiary contacts. We now test our model for different combinations of parameters m_r, m_s and m_t as listed in table:3.2. Here we consider $m_r = 1$ where as Toivonen et al. considered $m_r = 1$ with probability 0.95 and $m_r = 2$ with probability 0.05 for picking initial contact. For larger values of m_r in our model, the network growth is increasing drastically. Hence we have fixed the value of m_r to be 1. m_s and m_t are chosen a probability distributions U[0, k] where U[0, k] is uniform probability distribution on [0, k].

Parameter set	Initial cont.(m_r)	Secondary cont.(m_s)	Tertiary cont.(m_t)
1	1	U[0,4]	U[0,3]
2	1	U[0,3]	U[0,2]
3	1	U[0,2]	U[0,2]
4	1	U[0,2]	U[0,1]

Table 3.1: Above table showing parameters setting value for initial contact is 1, and secondary and tertiary are changed.

The tertiary attachment model is simulated with these different sets of parameters. We start with a seed network of 1 node and grow it upto 2000 nodes with the different combinations of parameters. The test results are presented in table:3.2.

Clearly the average clustering coefficient is largest for parameter set 1 and is decreasing for parameters sets 2,3 and 4 denoting formation of larger number of clusters for case 1. Similar is the case for average degree. And naturally the average path length is smallest in the case of parameter set 1 and is more for the other cases. As the network grows, these characteristics are plotted in table: 3.2, figures: 3.6, 3.73.8.

	$IC=1$ $SC=U[0,4]$, $TC=U[0,3]$	$IC=1$ $SC=U[0,3]$, $TC=U[0,2]$	$IC=1$ $SC=U[0,2]$, $TC=U[0,2]$	$IC=1$ $SC=U[0,2]$, $TC=U[0,1]$
Network size	2000	2000	2000	2000
Initial Connections	1999	1999	1999	1999
Secondary Connections	2148	2154	1551	1552
Tertiary Connections	3461	2197	2030	1732
Total no.of Connections	7608	6350	5580	5283
Total no.of triangles	6385	3832	3744	3272
Average degree	6.32	5.55	4.89	4.55
Avg. clust.coefficient	0.57	0.52	0.50	0.48
Average path length	4.63	5.60	5.79	6.24
No.of components	1	1	1	1
Max. degree	186	152	126	110

Table 3.2: Average statistics obtained for a network of size 2000 vertices when the tertiary attachment model is simulated with different sets of parameters as listed on each column.

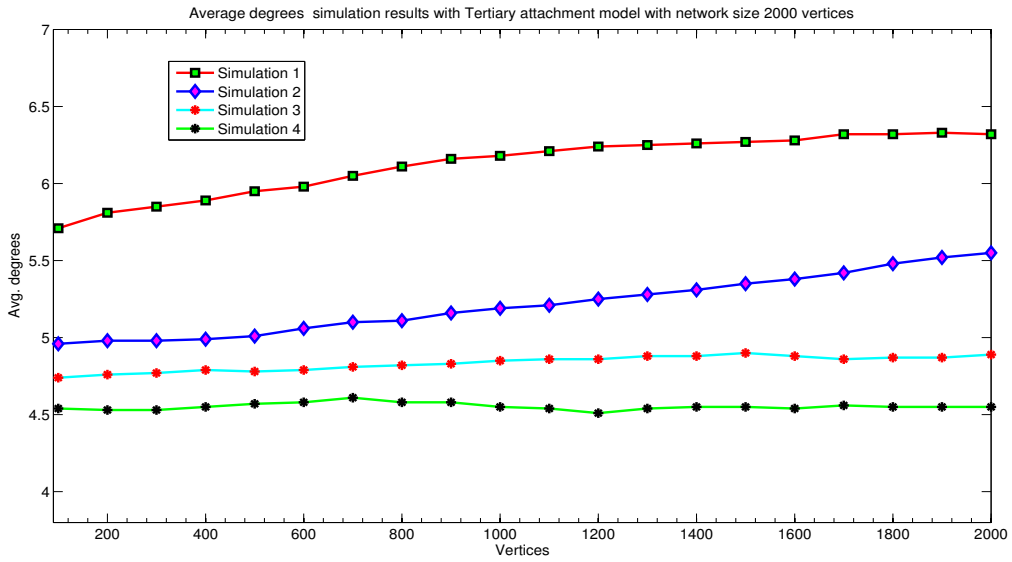

Figure 3.6: Tertiary attachment model is simulated with different parameter sets listed in table: 3.1 as the networks grows upto a size of 2000 vertices. In simulation 1 to 4 m_s and m_t values gradually decreased.

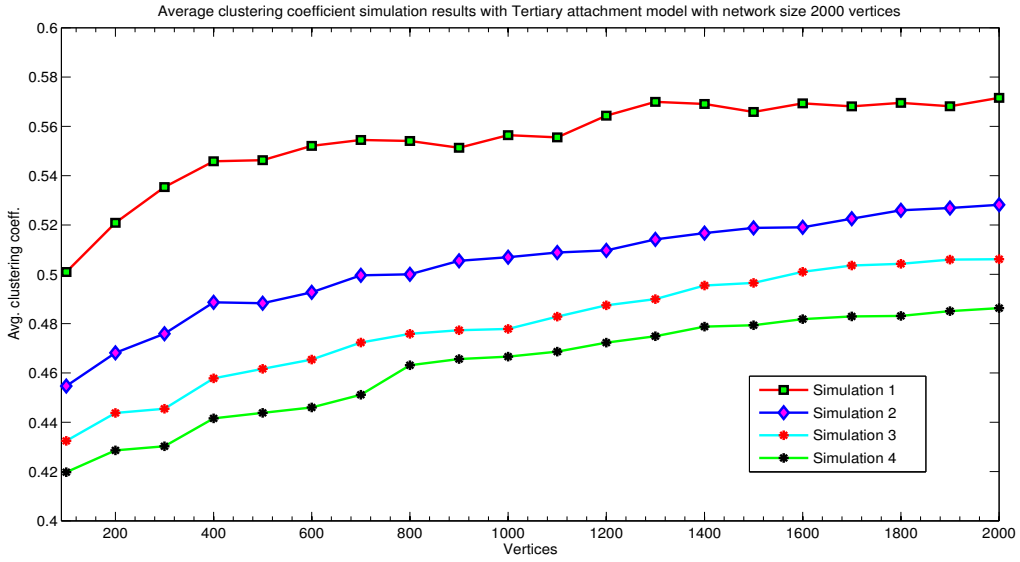

Figure 3.7: Comparison of average clustering coefficient as the networks grows upto size of 2000 vertices according to Tertiary attachment model, for the four different parameter sets.

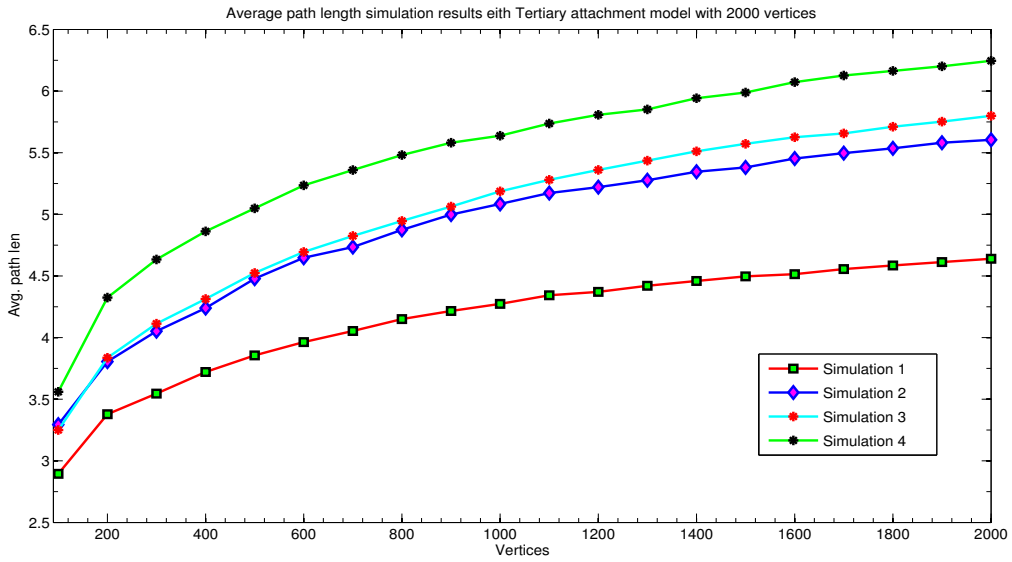

Figure 3.8: Plotting average path length for networks of size to 2000 vertices according to tertiary attachment model. for different parameters sets with $2^{nd} and 3^{nd}$ cases close to the a line.

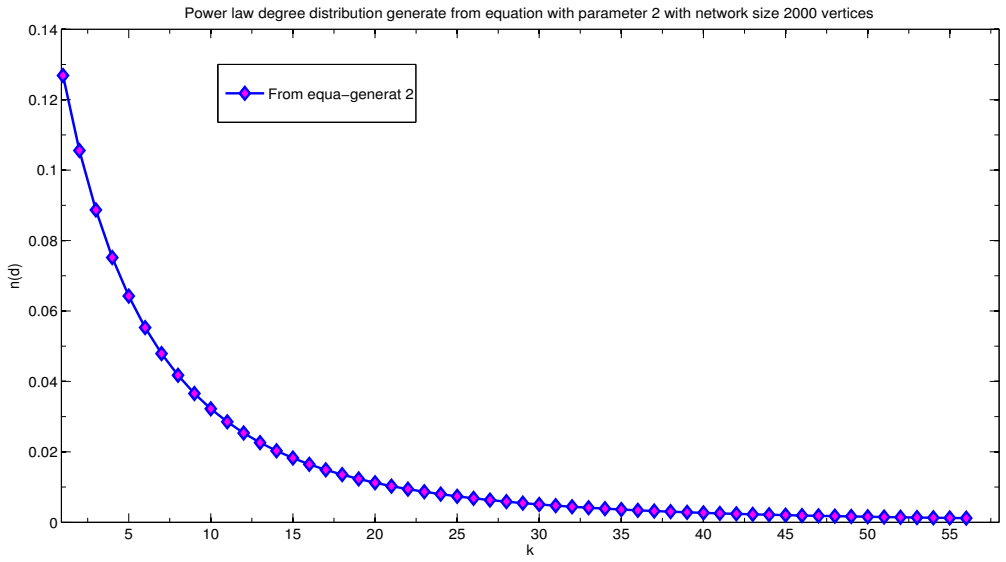

Figure 3.9: Simulation of the rate equation 3.7 for degree for $m_r = 1$, $m_s = \mathrm{U}[0,3]$ and $m_t = \mathrm{U}[0,2]$, rate equation, power law degree distribution with Tertiary attachment model for a network size as 2000 vertices.

3.9.1 Validation of the model

The tertiary attachment model is sought to be validated by plotting the theoretical curves obtained from the rate equations 3.7against the plots obtained by simulating

the model. The comparison is carried out for all the four different parameters sets.

The average clustering coefficient at a node v_i, $c_i(k_i)$ in the equation obtained from the rate equation 3.13 is plotted by substituting the average degree of m_r, m_s and m_t for particular distribution as given table: 3.1 and 3.3.

Parameter set	Average m_r	Average m_s	Average m_t
1	1	m_s=2	m_t=1.5
2	1	m_s=1.5	m_t=1
3	1	m_s=1	m_t=1
4	1	m_s=1	m_t=0.5

Table 3.3: Average values chosen for different parameter sets.

In the figures: 3.10, 3.11, 3.12 and 3.13, for the different parameters sets, the rate of change of average clustering coefficient at a node as the network grows from one node as seed to 2000 vertices is plotted, from the equation as well as the simulation of the corresponding model.

Visually, parameter set 2 and 3 seem to fit the model closest to the theoretical results. More importantly, for all the parameters sets, the plots obtained from the simulated model are highly correlated to the theoretical equation plots which validates the model. The power-law underlying the degree distribution is clearly visible in the plot figure: 3.14 for all the different parameters set.In which [1, 2, 3 and 4] are equally likely since $\cup[0,4]$ is the distribution taken from m_r, and average value of 2 is taken were for m_r to be substituted in the equation 3.12.

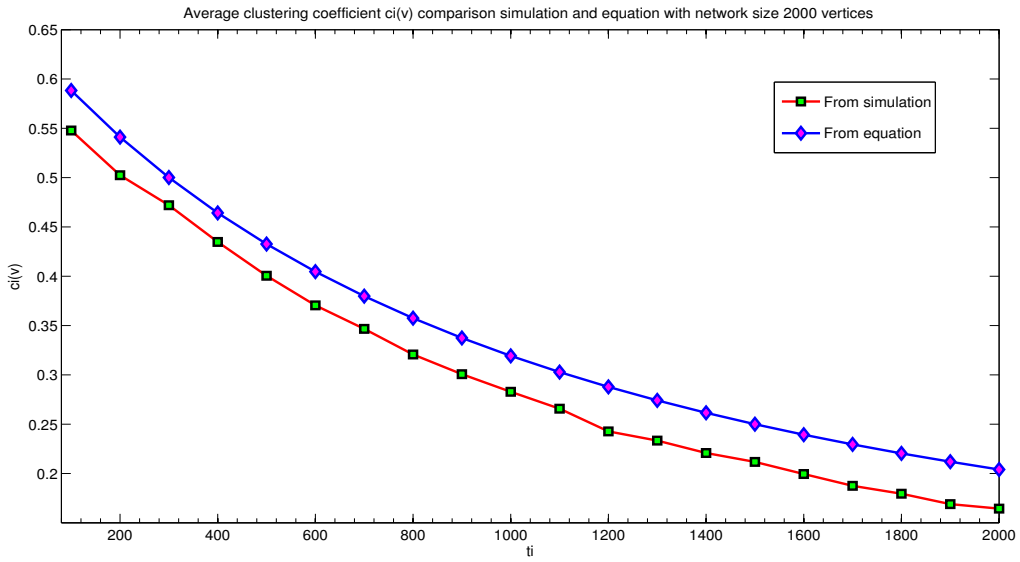

Figure 3.10: Clustering coefficient c(k) averaged over 10 iterations on a network size of 2000 vertices with calculated $m_r = 1$, $m_s =\sim U[0,2]$ and $m_t =\sim U[0,2]$.

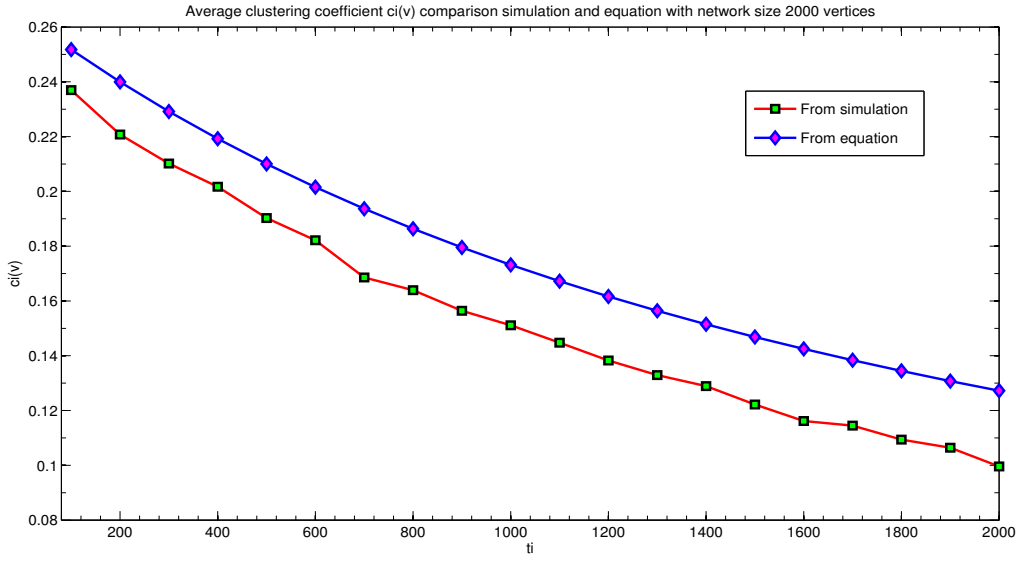

Figure 3.11: Clustering coefficient c(k) averaged over 10 iterations on a network size of 2000 vertices with the parameter $m_r = 1$, $m_s =\sim U[0,2]$ and $m_t =\sim U[0,1]$.

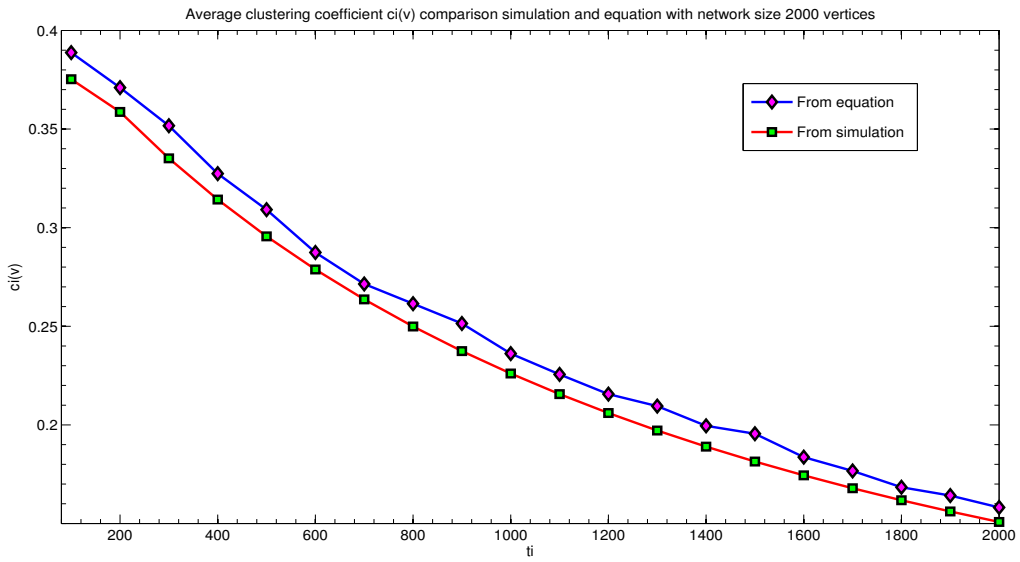

Figure 3.12: Clustering coefficient c(k) averaged with 10 iteration with network size 2000 vertices with different parameters see table: 3.2. Parameters observed with $m_r = 1$, $m_s =\sim U[0,2]$ and $m_t =\sim U[0,2]$ results satisfied small world property

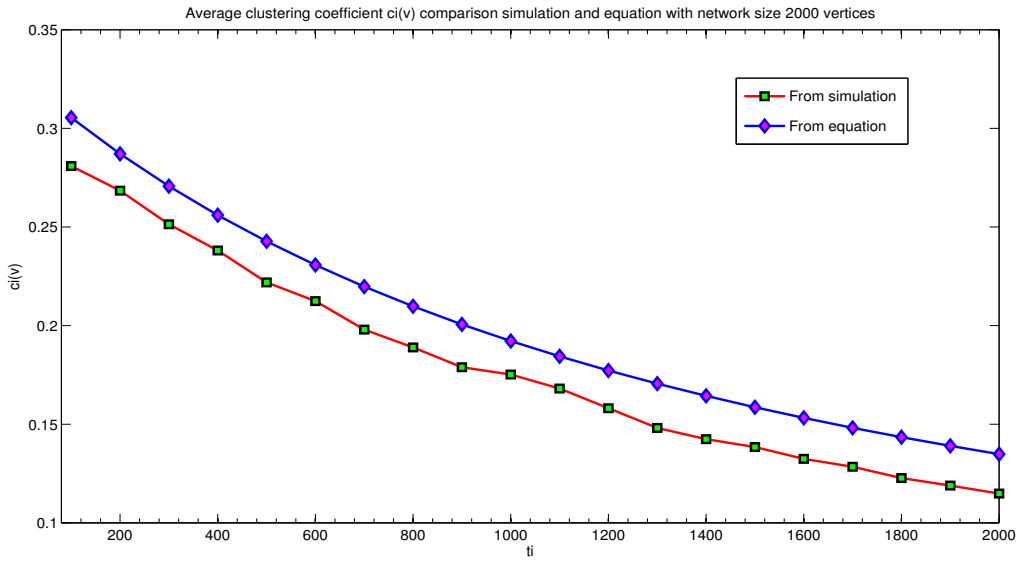

Figure 3.13: Clustering coefficient c(k) averaged with 10 iteration with network size 2000 vertices with different parameters see table: 3.2. Parameters observed with $m_r = 1$, $m_s =\sim U[0,2]$ and $m_t =\sim U[0,1]$ results satisfied small world property

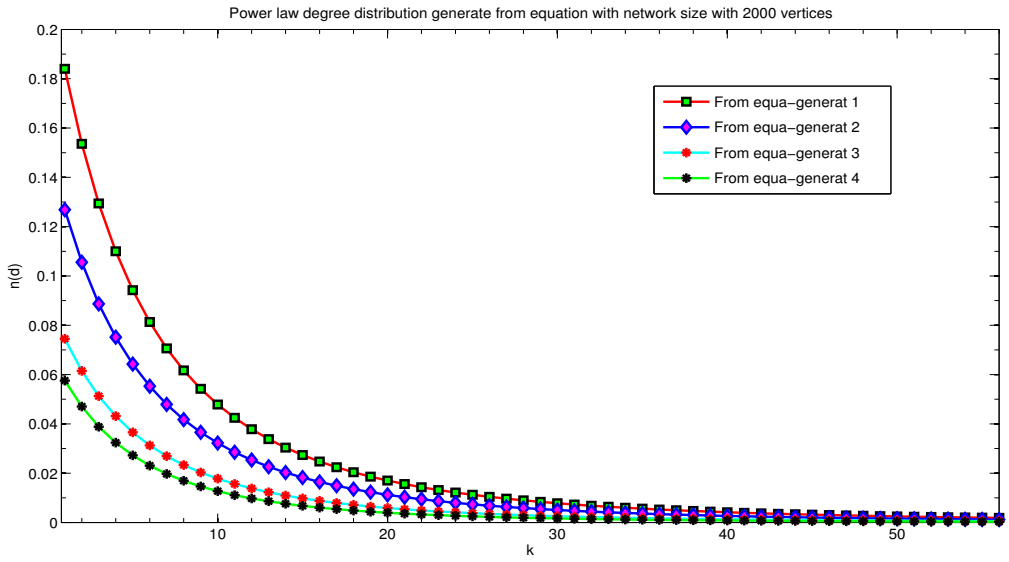

Figure 3.14: Simulation of rate equation 3.7 for degree distribution with the four different parameters.

3.10 Simulation of the model: Average statistics

The model is simulated $m_r = 1$, $m_s = U[0,3]$ and $m_t = U[0,2]$,where $U[0,k]$ is uniform distribution on $[0,k]$, by fitting to grow a network upto a size of $10,000$ vertices. Due to the stochastic nature of the model, we repeat each of the experiments 10 times and present average statistics of the networks thus obtained. Firstly the different types of connections that are made during the evolution of the network are computed. The number of edges formed due to primary, secondary and tertiary contacts as a function of number of nodes is shown in figure: 3.15. The figures show that the number of edges due to initial contact is less compared to those due secondary or tertiary contacts. This is a natural scenario in, for example, friendship network, new friendships are most often formed through tertiary or secondary and direct friendships are relatively less. In addition, edges due to tertiary are more compared to secondary because, we included the internal edges as part of tertiary connection.

The formation of triangles in our model is compared with that of Toivonen et al. and shown in figure: 3.16. Obviously, the proposed model produces more triangles compared to earlier model due to the introduction of tertiary contacts. Also

one can observe a higher slope of growth in number of triangles produced in excess compared to earlier model which leads to a complex network structure as network grows.

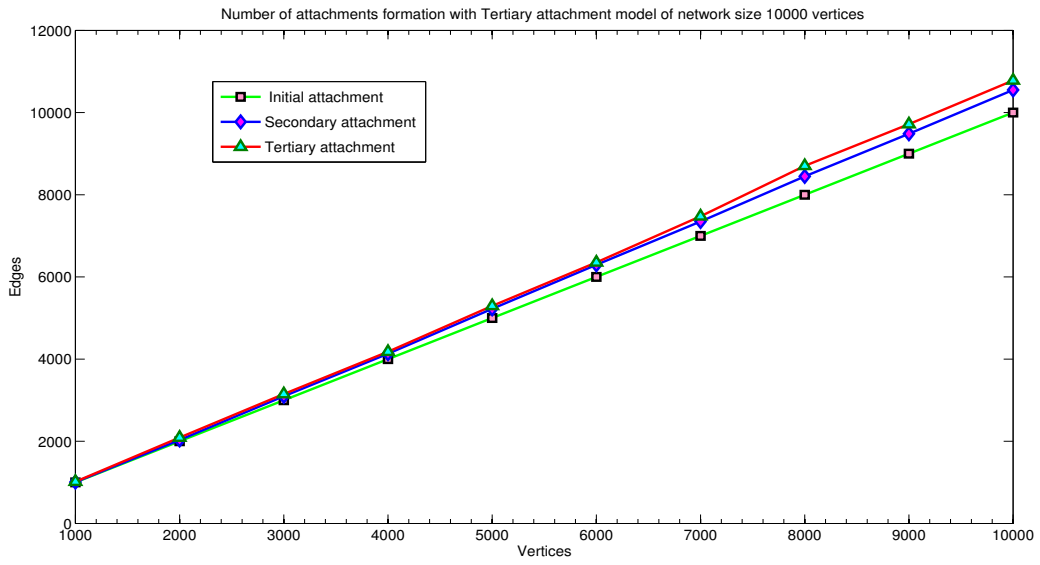

Figure 3.15: Number of Initial, Secondary and Tertiary connections as a function of node count for 10000 vertices. Also note that # initial nodes in the graph = #initial attachment, since $m_r = t$ with $m_r = 1$ uniformly. Edges due to tertiary connections larger than due to secondary and initial contacts.

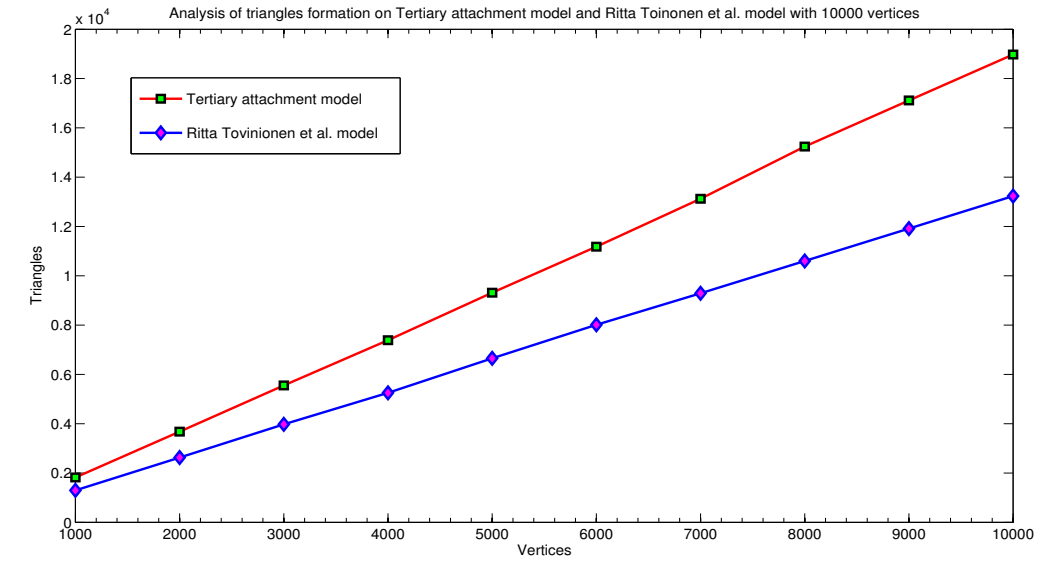

Figure 3.16: Comparison of triangle formation between tertiary attachment model and Toivonen model for 10000 vertices. Slope of the curve is increased from 1.326 for that of Toivonen et al. to 1.906 for tertiary attachment model due to the introduction of tertiary contacts.

3.10.1 Average degree

Average degree grows via three ways: firstly when a vertex directly links to v_i, secondly when initial contact vertex updates its friend of friend and thirdly when the new node connects to tertiary contact. In the social network graphs shown in figure: 3.17 for 2000 vertices and in figure: 3.18 for 10,000 vertices, one can observe a slow growth rate in average degree in tertiary attachment model as well as secondary attachment model of Toivonen et al. with slope of variation of total degree in the range $(10^{-6}, 10^{-4})$ clearly the trend of growth of average degree is similar in both the models.

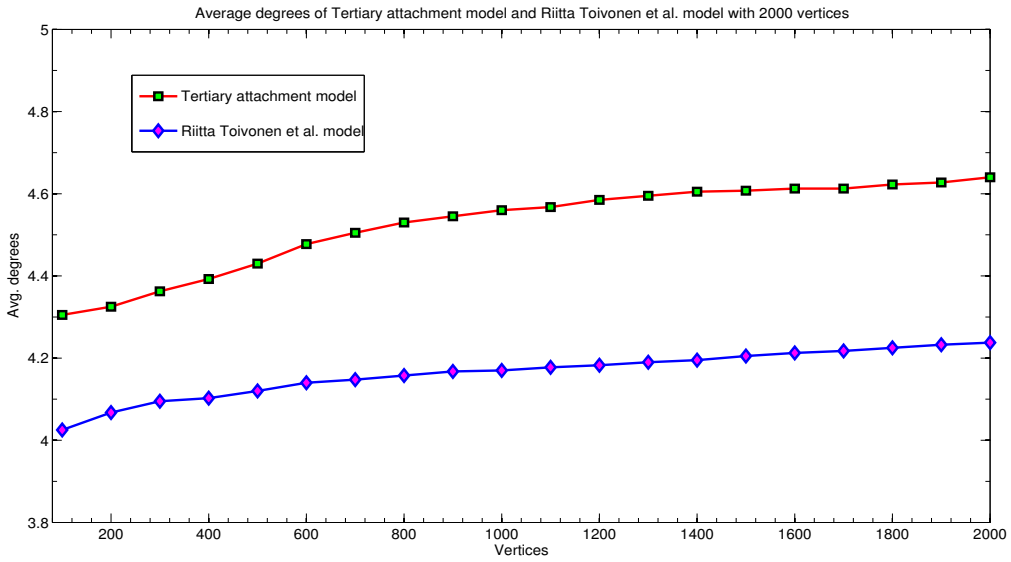

Figure 3.17: Average degree of networks of sizes varying in steps of 100 nodes grown to 2000 vertices according to Tertiary attachment model and Toivonen model. Tertiary attachment model shows slope varying in the range $(10^{-5}, 4*10^{-4})$ with an average value of 4.64 where as Toivonen model shows a slope varying in the range $(7*10^{-5}, 11.7*10^{-4})$ with an average value of 4.23.

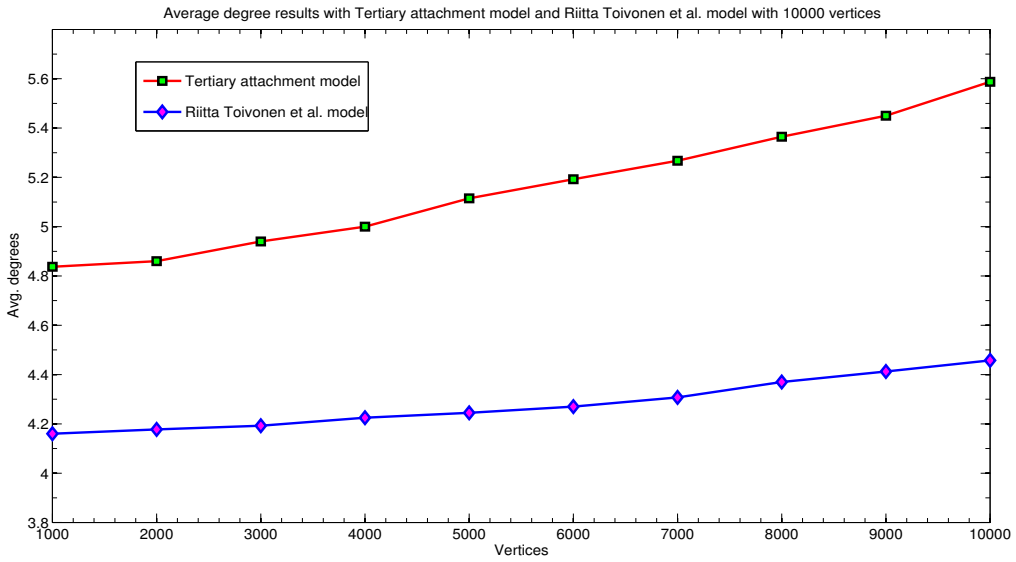

Figure 3.18: Comparison of average degree of networks of sizes varying from 1000 till 10,000 vertices grown using tertiary attachment model and Toivonen model. Tertiary attachment model shows slope varying in $(5.5*10^{-6}, 2.8*10^{-4})$ with an average value of 5.58, Toivonen model shows a slope varying in $(8.5*10^{-6}, 1.9*10^{-4})$ with an average value of 4.45.

3.10.2 Average clustering coefficient

Average clustering coefficient grows via the triangles formed in the process: firstly when the initial contact updates its friend of friend and secondly when new node connects to tertiary contact and thirdly when new node selects neighbour of initial contacts as secondary contact. The variation of average clustering coefficient as a function of number of nodes is presented in figure: 3.19 and figure: 3.20 for networks of sizes 2000 and 10,000 respectively. We can see that as the size of the networks increases the clustering coefficient of the tertiary attachment network is more when compared to that of network grows according to model of Toivonen et al. Also, it is significant that the two curves are highly correlated.

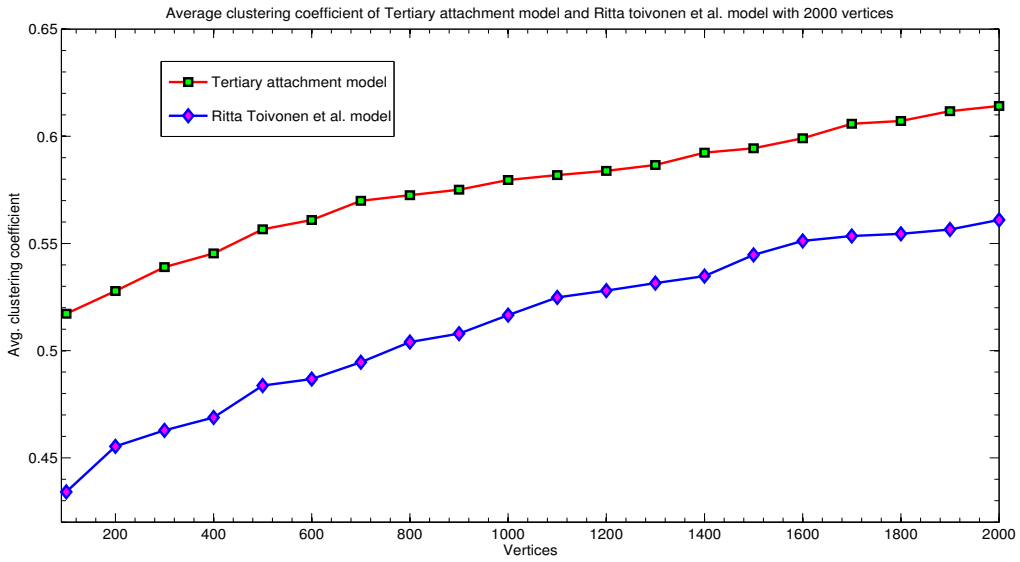

Figure 3.19: Average clustering coefficient values for our model and model of Toivonen et al. are plotted at different time stamps for a network of 2000 nodes. A mean value for clustering coefficient of 0.61 is obtained for tertiary attachment model and 0.56 for secondary attachment model.

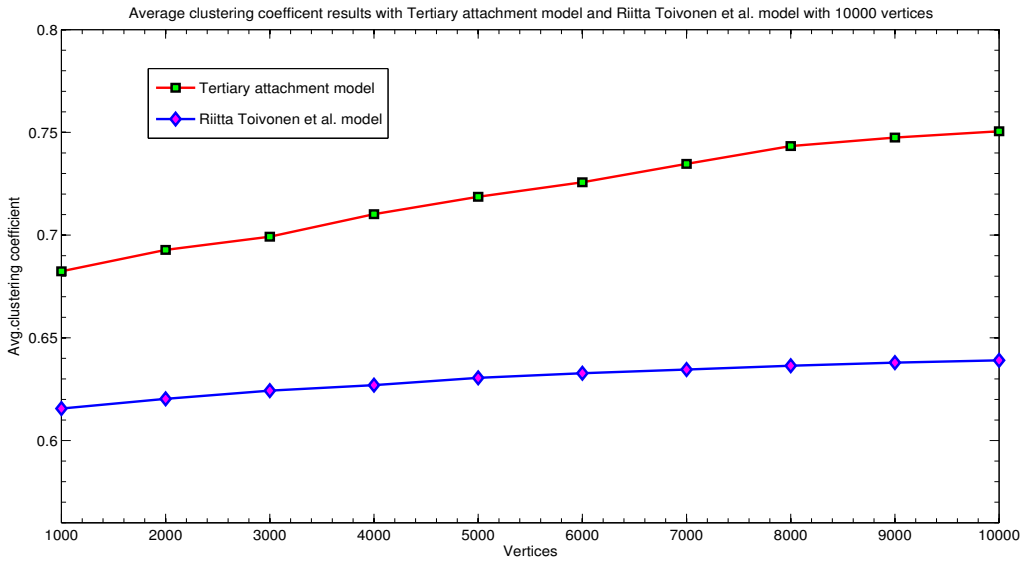

Figure 3.20: Average clustering coefficient comparison for our model and Toivonen model with a mean value at 0.75 and 0.639 respectively for a network of 10,000 nodes.

3.10.3 Average path length

Average path length is the average number of steps taken along the shortest path between any two nodes of a network. The shortest path between any two nodes is called $geodesic$. All geodesics between a given pair of nodes will have same length, by definition. If d_{ij} is the length of the shortest path between the nodes i and j of a network, then average path length of an undirected graph of N nodes $\sum_{i,j} d_{i,j}/N$. A comparison of average path length for our model and model of Toivonen et al. has been presented in figures: 3.21 and 3.22 for node counts 2000 and 10,000 respectively. Similar to Toivonen graph, the average path length has very less fluctuations and is almost independent of number of nodes in the network. This makes our model to realize as scale free network similar to earlier model. Since the vertices of average path length of a model is smaller compared to Toivonen, this property allows for speedy propagation of information compared to earlier model.

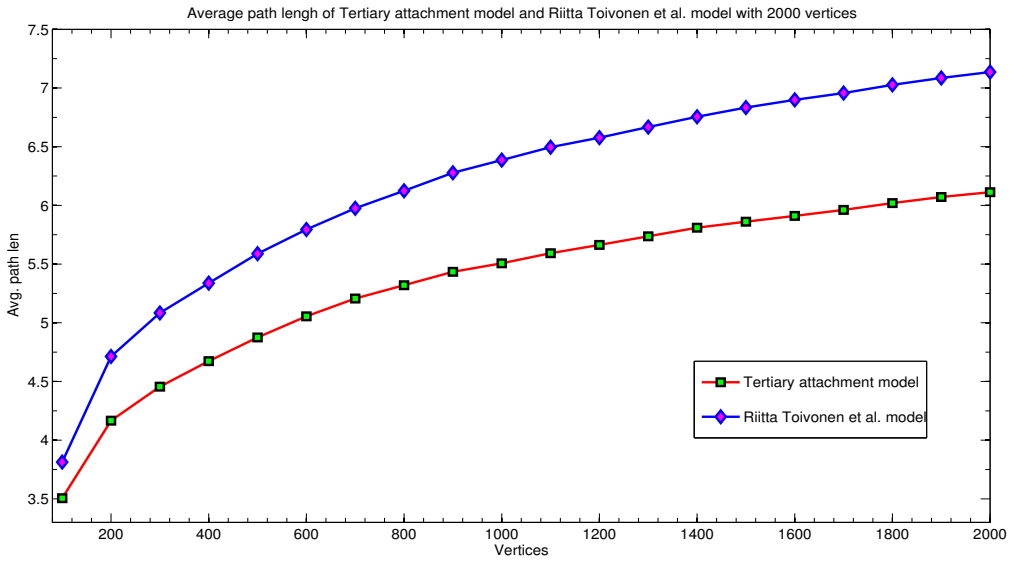

Figure 3.21: Average path length of networks of sizes varying one to 100 till 2000 vertices according to Tertiary attachment model and Toivonen model. Tertiary attachment model shows slope varying from $(2.3*10^{-4}, 1.9*10^{-3})$ with an average value of 6.11 where as Toivonen model shows a slope varying in $(6.8*10^{-4}, 5.5*10^{-3})$ with an average value of 7.13.

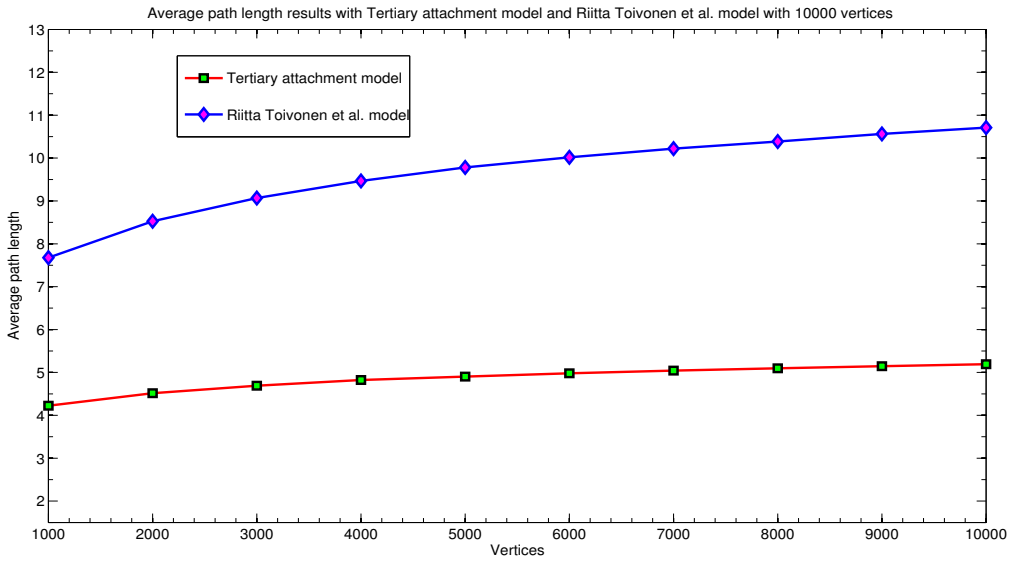

Figure 3.22: Average path length from 1000 to 10000 nodes. Tertiary attachment model shows slope varying in $(5.8 * 10^{-5}, 2.3 * 10^{-4})$ with an average value of 4.86 where as Toivonen model shows a slope varying in $(18.4 * 10^{-5}, 9.9 * 10^{-4})$ with an average value of 9.64.

3.10.4 Assortative mixing

Vertex degree-degree correlations is commonly measured using average nearest-neighbour degree spectrum $k_{nn}(k)$. If $k_{nn}(k)$ has a positive slope, high-degree vertices tend to be connected to other high-degree vertices called assortative mixing. Our tertiary attachment model exhibits assortative mixing property with higher degree vertices connecting to other high degree vertices.

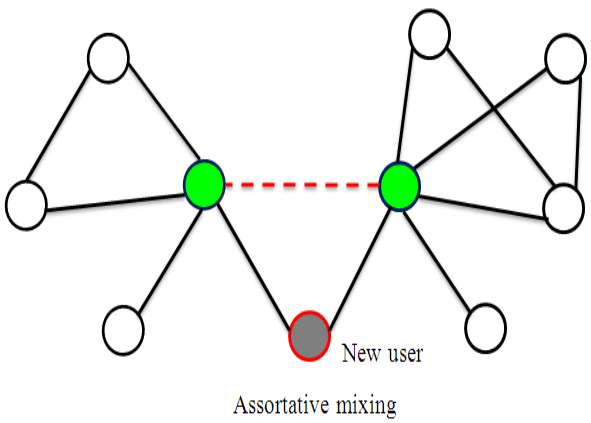

Figure 3.23: It is possible in tertiary attachment model, that interaction is established between two initial contacts which are high degree nodes of two communities interaction reflects asssortative mixing.

In figure.3.24 $k_{nn}(k)$ is plotted for different degree and it can be seen that tertiary attachment model has higher assortative mixing. When compared to that of secondary attachment model.

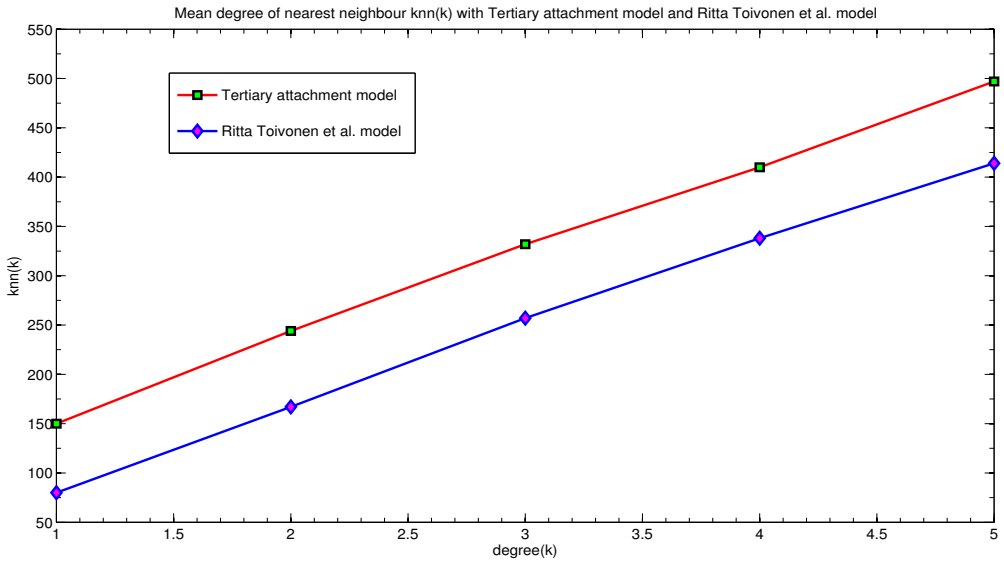

Figure 3.24: Pointwise average nearest neighbor degree of a node in a network of 2000 vertices for tertiary attachment model and Toivonen et al. model. and a correlation of 0.964 is found.

3.10.5 Degree distribution

The degree distribution of the network graph obtained by simulating the tertiary attachment model and Toivonen model for $10,000$ vertices show clearly the power law satisfied by both the distributions. The plot is shown in figure: 3.25.

3.11 Summary results

We initialize the network with 1 seed and fix parameters $m_r = 1$, $m_s = \sim \cup[0,3]$ and $m_t = \sim \cup[0,2]$. The model is simulated to produce networks of sizes 1000 to 10000 vertices. These networks are analyzed for the properties of average degree, average clustering coefficient and average path length which are tabulated in table: 3.4.

The statistics obtained by the tertiary attachment model are compared to that of

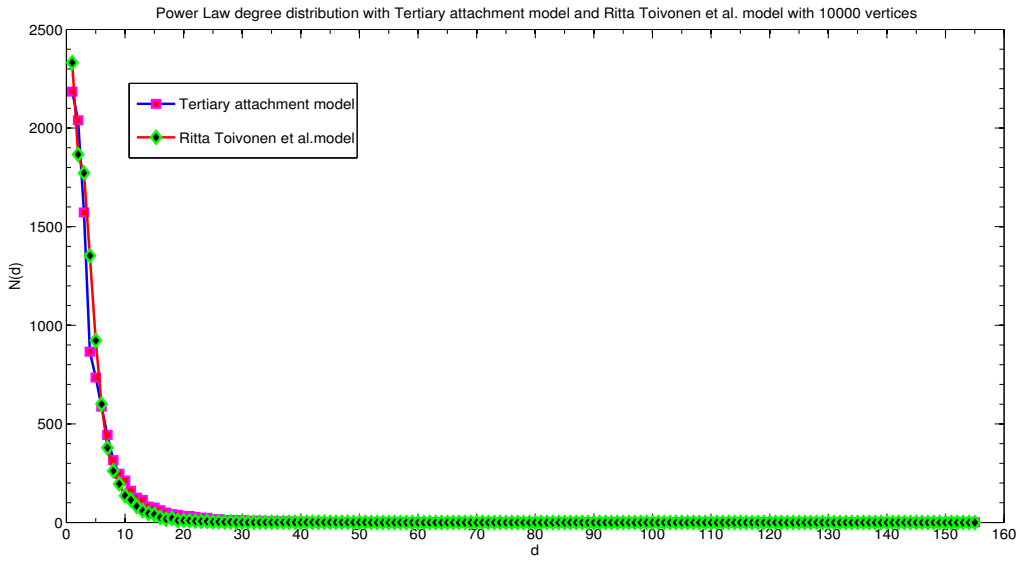

Figure 3.25: Above graph shows power law degree distributions obtained with Tertiary attachment model and Toivonen model with 10000 vertices with a correlation at 0.97.

Nodes	IC	SecCont	TerCont	Avg.Degree	Clust.Coeff	Avg.Path Len.
1000	999	1007	1016	4.85	0.687	4.22
2000	1999	2031	2091	4.88	0.688	4.51
3000	2999	3096	3149	4.92	0.698	4.69
4000	3999	4131	4179	4.98	0.707	4.82
5000	4999	5224	5295	5.12	0.717	4.90
6000	5999	6295	6353	5.18	0.726	4.98
7000	6999	7348	7476	5.22	0.737	5.04
8000	7999	8446	8703	5.31	0.748	5.09
9000	8999	9483	9718	5.38	0.755	5.14
10000	9999	10550	10781	5.42	0.758	5.19

Table 3.4: Average degree, average clustering coefficient and average path length of networks of sizes varying from 1000 to 10,000 vertices.

Toivonen et al. for graph sizes $N = 1000, 2000, 5000$ and $10,000$ and are tabulated in table: 3.5

We have simulated the models using some fixed parameter distributions and grown the graphs upto $10,000$ vertices for which average statistics is computed.

	Tertiary attachment model				Toivonen et al. model			
	Network	Network	Network	Network	Network	Network	Network	Network
Vertices / Nodes	1000	2000	5000	10,000	1000	2000	5000	10,000
Average degree	4.01	4.25	5.16	5.43	3.61	4.13	4.19	4.28
Average clustering coefficient	0.51	0.56	0.67	0.73	0.46	0.51	0.56	0.62
Average path length	4.64	5.34	5.54	5.19	5.79	6.17	7.68	9.64
No of components	1	1	1	1	1	1	1	1
Maximum degree	62	96	145	193	48	71	98	123

Table 3.5: Snapshot of results for graphs of sizes 1000, 2000, 5000 and 10,000 vertices with Tertiary attachment model and Toivonen model.

3.12 Discussion and conclusions

In this chapter we have proposed tertiary attachment model for social networks extending the attachment model of R.Toivonen *et al.* (2006). Both models have been developed based on the preferential attachment mechanism proposed by Barabasi and Albert (1999). The motivation behind the model is based on a realistic scenario:

Let us consider a real world problem where a person contacts a person in a research group for help and suppose that he did not get adequate support from this initial contact or his neighbors. But he may get required support from a friend of friend of his initial contact. Then the only way the new person could get help is that his primary contact has to update his friend of friend for supporting the new contact and introduce him to his friend of friend.

In these technology days, let us consider a real world problem where a person A contacting a person B through mobile to visit a tourist place for his own purpose. suppose B or B's neighbors do not have the information required by A, suppose B's friend of friend C possesses the information required by A, then one way A can get help is if B updates friend of friend list (makes contact with C) thus facilitating A to make contact with C.

As an extension of our Tertiary attachment model, one can go for the quaternary contact. ie; the new node makes a contact with the friend of friend of friend of friend of initial contact. But the problem is this will lead to a complex network which is densely connected. This may not reflect the real world community structure in the grown network. Also practically reliability issues arise while making a contact with the friend of friend of friend of initial contact. These connections may not be trust worthy in the real world cases. Keeping these two issues in mind, we have extended only upto tertiary contact and not further beyond.

The average statistics of characteristics of social network like degree, clustering coefficient, assortative mixing etc.. have been empirically estimated for the proposed model by simulating the model for various sizes of networks.

It is presented that the average degree, clustering coefficient and maximum degree increase compared to the earlier model where as the average path length decreases. Moreover, a significant correlation between all the corresponding plots has been found.

The tertiary attachment model is also simulated for four different parameter sets and for all simulations there is a high correlation found between the corresponding plots. Further, theoretically, rate equations for average degree of a node and clustering coefficient at a node have been proposed. The model simulations have been found to nicely fit the theoretical estimates, further validating the proposed model.

Investigation of the model for existing benchmark datasets like GP and Facebook is carried out in the next chapter.

CHAPTER 4

TERTIARY ATTACHMENT IN REAL WORLD SOCIAL NETWORKS

4.1 Introduction

In this chapter we present the validity of our model by applying to benchmark datasets. For contrast, we have considered two different social networks. One is an online social network and another is an academic social network. These two social networks have distinguishable properties as one can observe. In online social networks, preferences and interactions will change from time to time very randomly and fluctuations are very high, where as in research and academic collaborations, interactions are mostly fixed and may not change with time drastically as compared to online social networks. Hence a study of the validity of our model to these kind of extreme cases will strengthen the applicability of our model to most of the real world datasets.

We described Tertiary attachment model (TA) which is an extension of secondary attachment model of R.Toivonen *et al.* (2006) in the last chapter. Here we explore the applicability of our proposed model in simulation and prediction of real world networks. We choose Genetic Programming (GP), a collaboration network and Facebook (2006-2008)(FB), a friendship network for our study since these are available with time stamp. We design experiments in which secondary and tertiary connections can be extracted from GP and FB. Additionally, we define different types of triads T1, T2, T3 and T4 based on secondary and tertiary connections, which we show to be significantly present in both GP and FB. This analysis shows the predictive power of the model, to predict the potential links that form in the future. Such an analysis has not been done for GP and FB in the literature to the best of our knowledge.

Further, we carry out simulations of the rate equations derived for the TA model, with different parameter distributions and match the rate of change of degree and clustering coefficient for GP and FB. For completeness sake we carry out the routine statistical analysis by plotting the growth of average degree, clustering coefficent and average pathlength and component analysis as a function of time (year/month/day) and compare with the average statistics of the network simulated by the model.

4.2 Literature review on collaboration networks

Newman (2001) built scientific collaboration network based on research interaction at Santa Fe institute and included year of publication. They empirically established that the probability of scientific collaboration increases with the number of common collaborations.

Tomassini and Luthi (2007); Luthi *et al.* (2007) conducted analysis of genetic programming (GP) co-authorship network which is part of the scientific co-authorship network, network grown according to preferential attachment with analysis of statistics of average degree, clustering coefficient, average path length, author collaboration and components in GP network.

Tang *et al.* (2008) present the design and experimental study of academic social networks with different datasets and they justify new metrics to look at award recipients in the computer science domain, like ranking a most influential author in the whole network.

Liu *et al.* (2012) propose a new evolution model of collaboration network based on the scale-free network, in which nodes are taken to be authors with some attributes. They compared the simulation with empirical data from CiteseerX database and the Motif emerging model to show that the statistical characteristics and evolution characteristics of their model conforms to real world datasets. On similar lines, we plan to analyze GP dataset with reference to our proposed tertiary attachment model.

4.3 Real-world datasets

There are very few benchmark datasets available with time stamp. We choose the collaboration network called $Genetic\ Programming$ referred to as GP dataset (Tomassini and Luthi, 2007; Luthi *et al.*, 2007). The GP bibliography created and maintained by W.B. Langdon and by S. Gustafson, is a database that contains almost all the papers published in the GP field since its inception between 1986 to 2006. We consider $Facebook$ users dataset which is basically a friendship network. The facebook data set is given with date of the link formation and the GP data set is available with the authors, year of publication of a paper. The choice of these datasets has been made because the time stamp of the edges is available in these datasets.

	Duration	No. of authors / Users	No.of edges
GP authors	1986-to-2006	2809	5850
Facebook users	Sept-06 to Dec-06	8620	20545
Facebook users	Jan-07 to Dec-07	24968	120201
Facebook users	Jan-08 to Dec-08	54161	458461

Table 4.1: Datasets of GP authors and Facebook users chosen.

4.4 Analysis of Genetic programming (GP) dataset

We carry out a detailed analysis of GP data set from the perspective of the tertiary attachment model. We would like to find out if two authors A and B collaborate in a certain year and say, B and C have a publication at a later year, is it true that there is a potential collaboration bound to happen between A and C. How many such collaborations happen? These links, if they exist, can be called as secondary attachments. Similarly do higher order links happen? For example, if C and another author D have a latest paper, is it true that there are links between A and D, a tertiary link, along with a collaboration between either A and C or B and D ?

We investigate these questions in this section by extracting different types of attachments and triads that exist in the GP data set. Analysis regarding the basic measures like average degree, clustering coefficient and path length is carried out by Tomassini et al. Also, analysis of a slightly different kind regarding connections formed

during preferential attachment is done by Tomassini and Luthi (2007). We perform a deeper analysis from the perspective of triads being formed as well as statistics of the component in this chapter.

First, we give the basic view of the data set. The details of GP dataset, that is the number of vertices and edges that get added each year to the network is given in the table: 4.2. A node in the GP network indicates an author who has atleast one bibliographic entry (an edge exists between this author and who have coauthored one or more papers or coedited atleast one book or proceeding). There are 2809 authors in total in the network from the years 1986 to 2006.

GP dataset contains a 106-author paper in 2005 which affects the analysis. Hence we exclude the paper from dataset. The dataset does not contain any single author nodes (isolated nodes).

	Input edges and authors		Cumulative author count and collaboration	
Years	Edges	No.of authors	Authors cumulative	collaboration
1986	1	2	2	1
1987	5	7	9	6
1990	1	2	11	7
1991	1	1	12	8
1992	12	15	27	19
1993	21	17	44	35
1994	73	71	115	99
1995	200	119	234	279
1996	347	176	410	562
1997	361	159	569	807
1998	510	219	788	1189
1999	567	222	1010	1604
2000	738	192	1202	2087
2001	629	223	1425	2516
2002	1001	302	1727	3275
2003	782	232	1959	3810
2004	915	311	2270	4483
2005	981	302	2572	5219
2006	890	238	2809	5850

Table 4.2: The information of new authors joining the GP networks each year resulting in increased collaborations is given.

4.4.1 Types of attachments

If a new node enters the graph and connects to an existing node(s) then the attachment is called as primary attachment. In other words, if a new new author publishes a paper in collaboration with an existing author, then the existing author is considered as primary contact and the connection established as primary attachment as seen in figure: 4.1 (G1, G2 and G3). If the new node connects to any one of the neighbors of primary contact, these connections are considered as secondary attachment (G5 and G6). If the new node connects to any one of the neighbors of neighbors of primary contact, they are considered as tertiary attachment (G9).

The figure: 4.1 gives different subgraphs that are extracted from GP data set. The nodes are labeled with author-id and the edge labeled with the year of publication between the two authors.

Let us now view the figure G5 closely: Note that the author 100 has two papers with 508 and 1403 in the year 1993. We see that 508 and 1403 do establish collaboration and publish a paper in 1994 proving the existence of secondary attachment. Similarly, in graph G6, nodes 78 and 105 published in 1990, authors 78 and 1529 have a paper in 1991. Then much in two years, in 1993, a friend-of-friend connection forms between 105 and 1529.

Now to observe tertiary attachments, see graph G9. The central node 78 has earlier collaborations with both 105 (in year 1990) and 995 (in 1995). Now a new author 2395 joins the network with a new collaboration with a multiple-author paper among 105, 995 and 2395. We term the edge between 105 and 2395 as a tertiary attachment and a internal edge between 105 and 995 is a secondary attachment.

The graph G7 shows a multiple-author paper published among 5 authors.

A plot of number of primary, secondary and tertiary connections as a function of year is presented in figure: 4.2. This shows that all the three types of contacts are increasing almost linearly from the year 1993 onwards before which there is no considerable change.

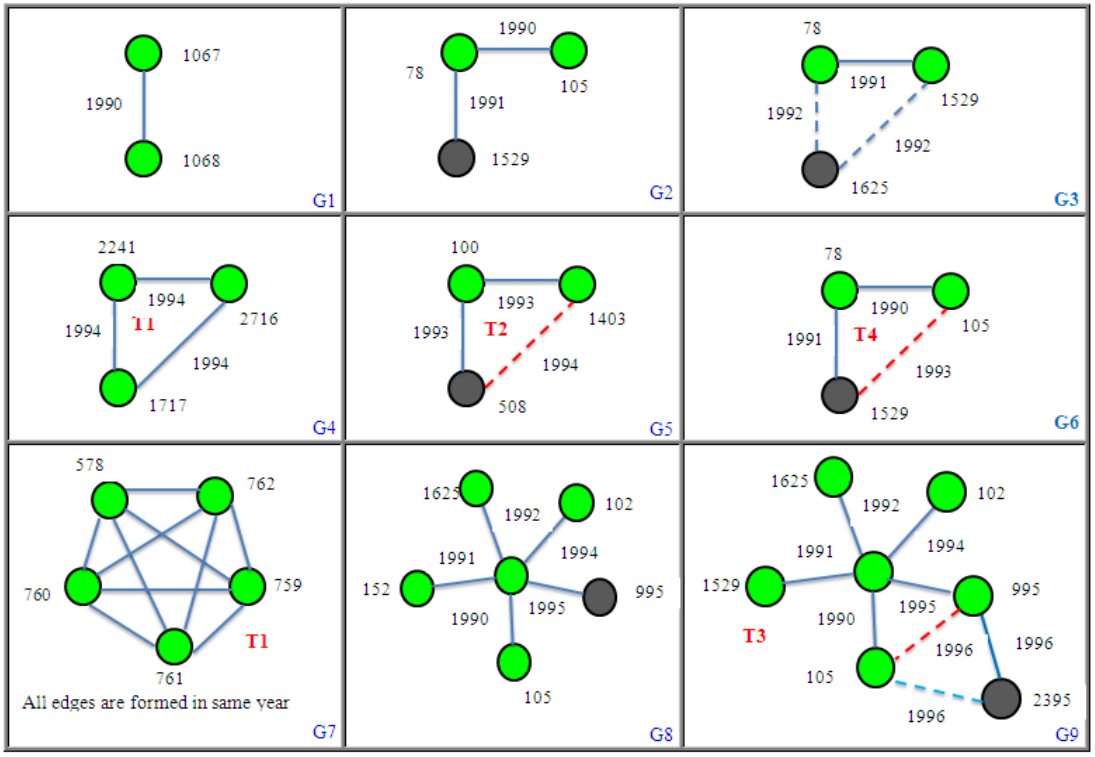

Figure 4.1: Different types of connections extracted from GP dataset are depicted here. The numbers on node represent author id and the year on the edge denotes the year of publication. Black node denotes the new node joining the network and edge being formed as the dotted edge

4.4.2 Collaboration

Collaboration is about working together to achieve a goal, process where two or more people or organizations work together to realize shared goals, it is well known that most of new collaborations are initiated through a common acquaintance. And if it is a like a chain of connections, the friend of friend introducing the new researchers to an acquaintance and it carries forward. It is also common experience in research groups that it is not easy to build new collaborations with totally unknown research groups.

Collaborations from the data set are extracted as multiple edges between two nodes with different year labels, where as an edge is counted only once. Types of collaborations can be summarized as follows: Given a seed network of authors, when a new author(s) joins the network, she joins into a collaboration with a friend among the existing authors (nodes), the figure: 4.1 depicts different types of collaborations that are possible in such a scenario.

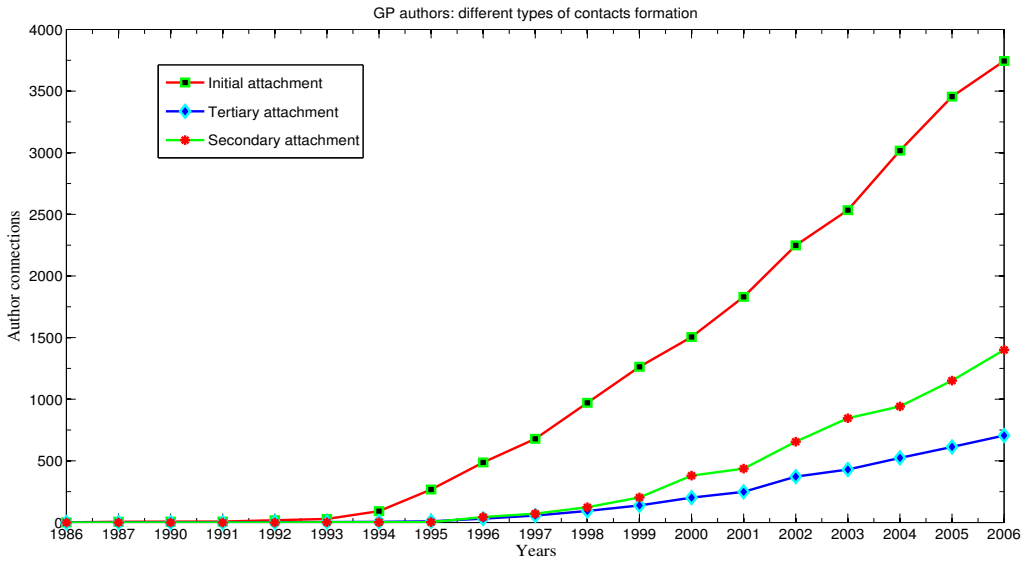

Figure 4.2: Plot of initial, secondary and tertiary attachments as a function of year for GP dataset from the year 1986 to 2006. By observing result from the above figure, the initial author collaboration contacts seem higher than other types of attachments. Also presence of secondary and tertiary attachments is seen to be significant amounting to about 40% of the total connections .

Figure: 4.3 is a snapshot of a subgraph of 13 nodes of the GP dataset with year labels depicting the different kinds of connections seen. For example, the edge connecting nodes 7 and 8 shows two collaborations. Figure:4.4 shows a subgraph of 200 authors from GP dataset in which a 12-author paper is clearly seen.

A comparison of node count to the edge count ie; a comparison of collaborations and authors count is presented in figure: 4.5. Collaborations being much more compared to the authors count shows that the GP dataset has more mutual collaborations. A comparison of new authors to the existing authors for GP dataset is presented in figure: 4.6.

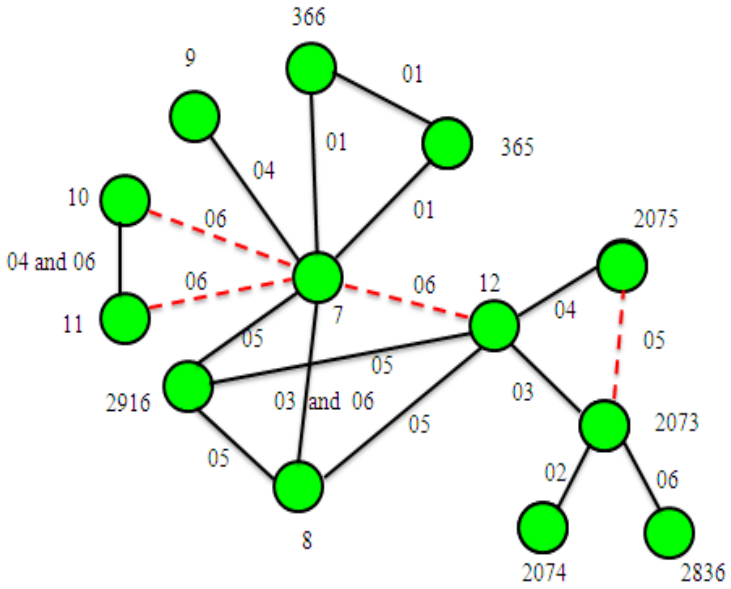

Figure 4.3: A snapshot of a subgraph of the GP dataset. The node id's are given along with the year of collaboration along the edge.

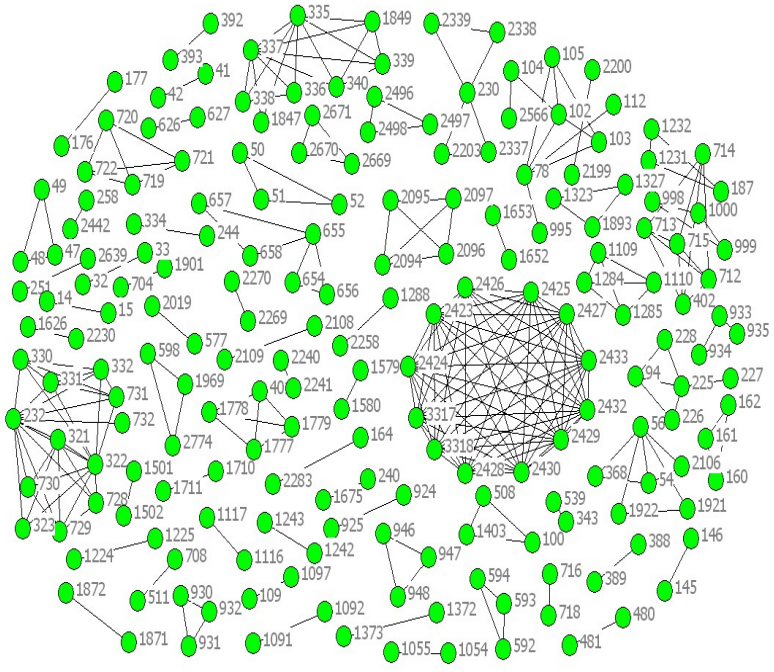

Figure 4.4: Growing process of academic social network GP with 200 authors from the year 1986 to 1996. A component of size 12 is clearly seen in the network and other dense as well as sparse interactions are seen.

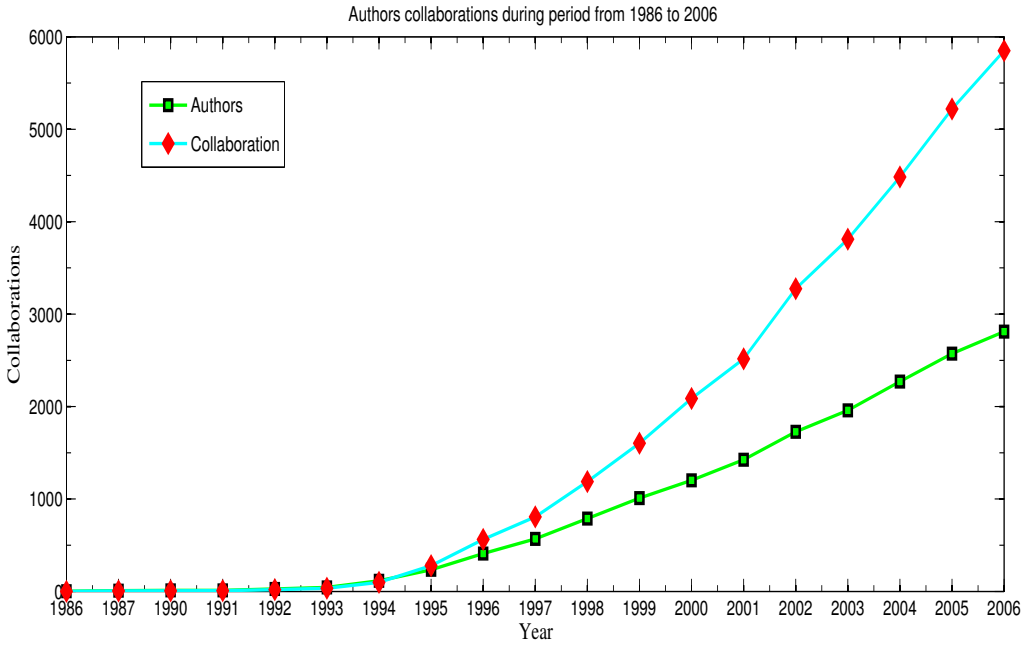

Figure 4.5: The authors joining at a linear rate with number of collaborations growing more steeply.

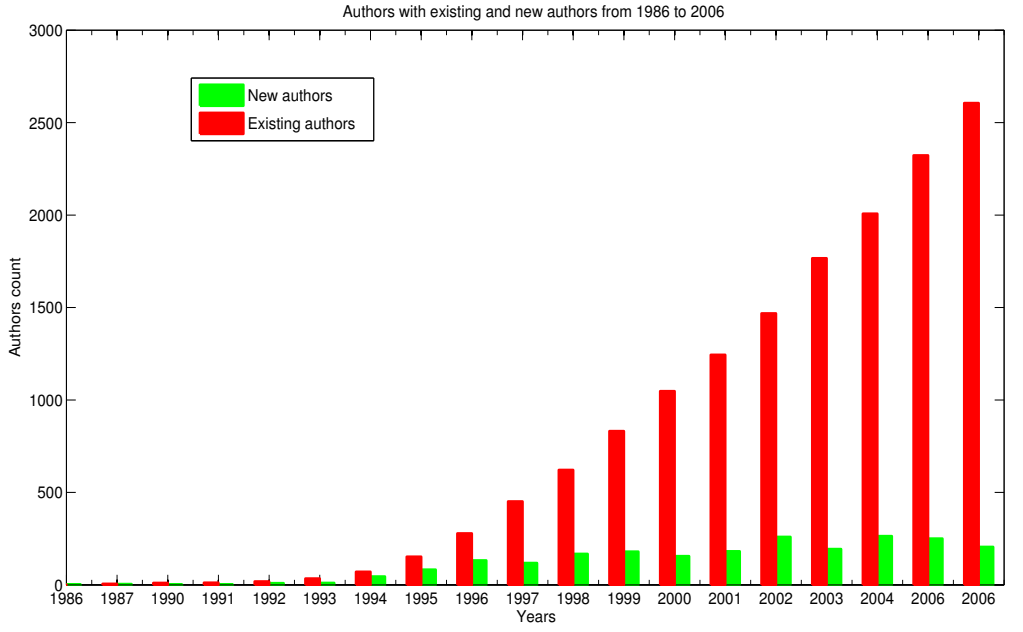

Figure 4.6: Observe that new authors who are joining are less as compared to existing authors.

4.4.3 Triad formation

Here we have considered four ways of formation of edges and in turn triangles. All the three edges may form in the same year (T1). Two edges may form in the same year and the third one may form in a different year (T2). One connection may form in one year and other two attachments may form in same year but different from the previous year (T3). All the three edges may form in different years (T4). These triangles are extracted from GP data set and depicted in figure: 4.1.

Coming to our model, triangle formation is via three processes: A new node may connect to any two nodes which are already in contact (G5) ie; the new node is selecting two connected nodes as primary contacts. Second mode is during the selection of secondary contacts (G6) ie; while selecting any of the contacts of primary contacts to form a triangle. Third mode is in the formation of the tertiary contacts (G9) ie; while selecting any of the neighbor of neighbor of primary contacts to form triangles.

In figures: 4.1 and 4.3 the actual node id's are given with the year of collaboration along the edges. The different types of triangles, T_1, T_2, T_3 and T_4 are also marked in figure: 4.1.

The number of triangles of different types in GP dataset present from 1986 upto 2006 the end of year is presented in tables: 4.3 and 4.4.

It can be seen that, multiple author papers (T_1) exceed the other type of collaborations. New collaborations forming between friend of friend (T_2) are also considerable. A higher order collaborations of the type T_3 and T_4 amount to 11 % out of the total triangles which is also significant.

Years	Triangle-T1	Triangle-T2	Triangle-T3	Triangle-T4
1986	0	0	0	0
1987	1	0	0	0
1990	1	0	0	0
1991	1	0	0	0
1992	2	1	0	0
1993	4	5	0	1
1994	27	6	0	1
1995	296	11	1	1
1996	511	62	14	2
1997	673	101	40	2
1998	1049	173	70	16
1999	1313	257	120	27
2000	1808	473	223	76
2001	2145	545	250	97
2002	3197	813	401	126
2003	3931	1117	462	198
2004	4472	1244	545	219
2005	5725	1545	627	281
2006	6785	1817	762	352

Table 4.3: Growth of triangles T_1, T_2, T_3 and T_4 in GP network.

	Academic network			Triangles formation			
Years	Initial attachment	Sec. attach.	Ter.attach.	T1	T2	T3	T4
4 years	23	6	1	1	0	0	0
8 years	376	63	62	296	11	1	1
12 years	1274	293	337	1313	257	120	27
16 years	2630	858	1118	3931	1117	462	198

Table 4.4: A cumulative assessment of different types of edges and triangles in GP over a step size of 4 years.

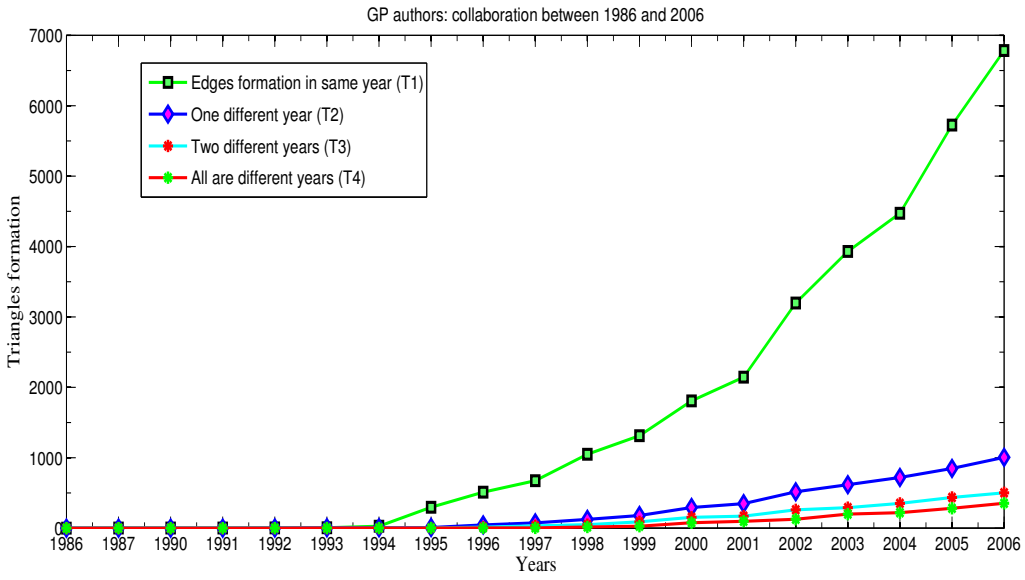

Figure 4.7: Clearly the number of triangles $T1$ which denotes multiple-author paper far exceeds the other types of triangles.

4.4.4 Average network statistics of GP

For completeness purpose, we present here some of the standard measurements for GP data set and we collate the same given in Tomassini and Luthi (2007) and found them to be matching. The average degree, average clustering coefficient and average path length for GP authors dataset are calculated and presented in table: 4.5. We have observed some of the papers have 20 authors. We found that there exists one paper with more than 100 authors, (106 authors, year 2005) which is drastically affecting the whole graph and causing an enormous increment to the values of average path length and clustering coefficient. We have excluded that paper from our dataset.

Years	Average degree	Average clus.coeff	Average path length
1986	1.00	0.00	1.00
1987	1.33	0.33	0.16
1990	1.27	0.27	0.12
1991	1.33	0.25	0.15
1992	1.41	0.28	0.07
1993	1.59	0.32	0.04
1994	1.72	0.41	0.02
1995	2.38	0.53	0.01
1996	2.74	0.56	0.01
1997	2.84	0.59	0.03
1998	3.02	0.60	0.09
1999	3.18	0.61	0.23
2000	3.47	0.62	0.27
2001	3.53	0.61	0.37
2002	3.79	0.63	0.37
2003	3.89	0.64	0.45
2004	3.95	0.65	0.55
2005	4.06	0.65	0.61
2006	4.16	0.66	0.63

Table 4.5: Year wise average degree, average clustering coefficient and average path length for the GP network.

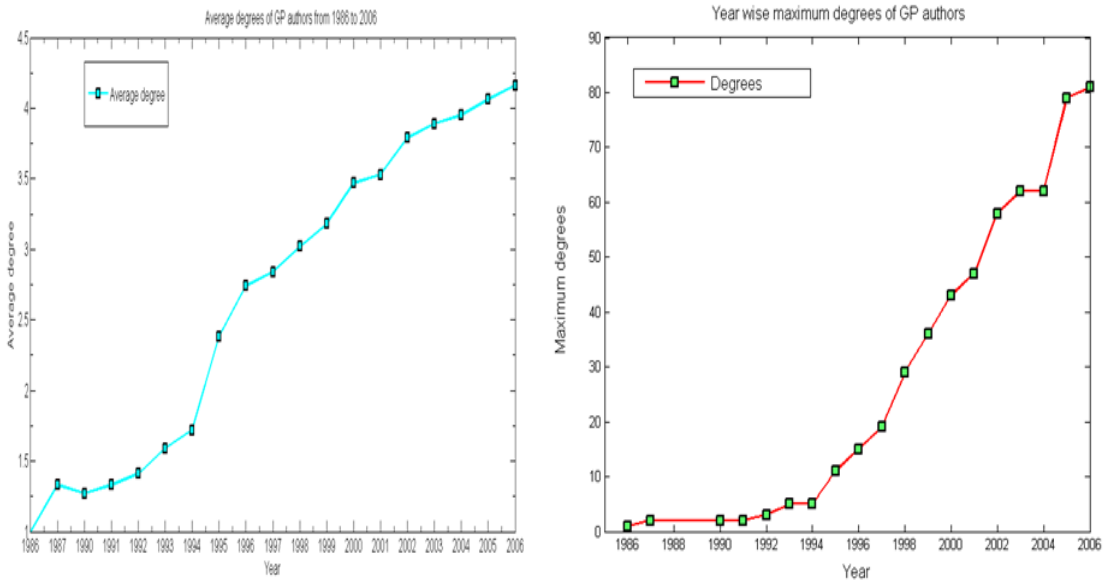

Figure 4.8: Above graph shows year wise average and maximum cumulative degrees calculated in GP network. It can be seen that during 1986 to 1994 there is a slow author growth rate after which there is higher growth rate.

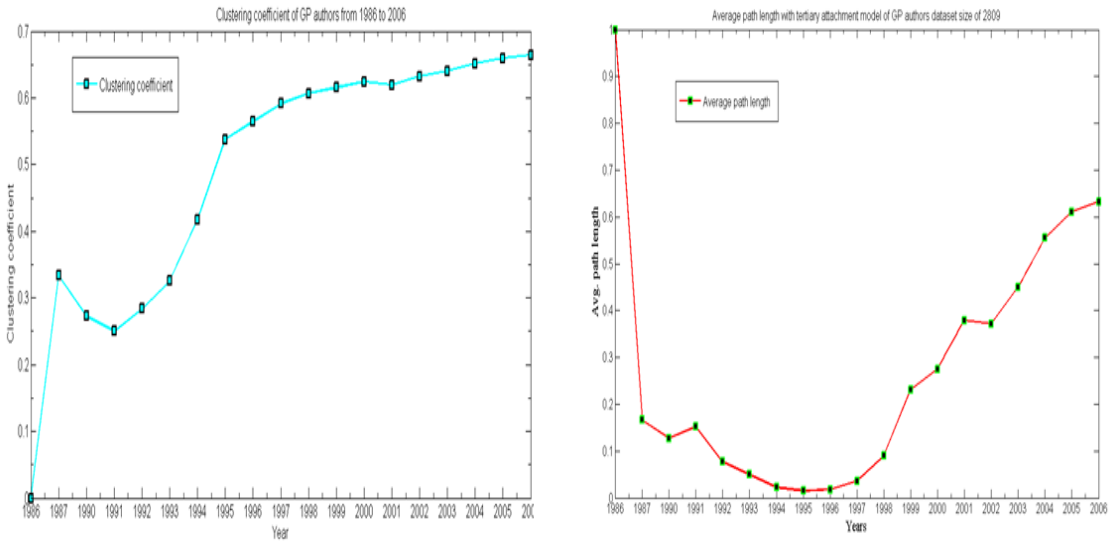

Figure 4.9: An unstable growth in clustering coefficent seen between 1986 to 1993 after which there is a steady increase in the value upto 2006. See in figure on the right a corresponding dip in average path length between 1992 to 1998 after which the value keeps increasing

4.4.5 Analysis of the largest component

Tertiary attachment model simulates the growth of a network as a single component. Hence, we study and analyze the largest component in GP dataset. A component is a maximal connected subgraph within a graph. Year wise increase in the number of components and the size of largest component in addition to maximum degree are presented in table: 4.6 for GP dataset. As one can see, there are less number of components with large degree nodes in GP data and there are more components with lesser degree.

Figure: 4.10 shows snapshot of components in GP dataset and figure: 4.11 confirms the result obtained by Tomassini and Luthi (2007); Luthi *et al.* (2007). We find that GP has largest component of size 1022 while the rest of the components are of very small sizes joining between 2-20, the details are given in the table:4.6.

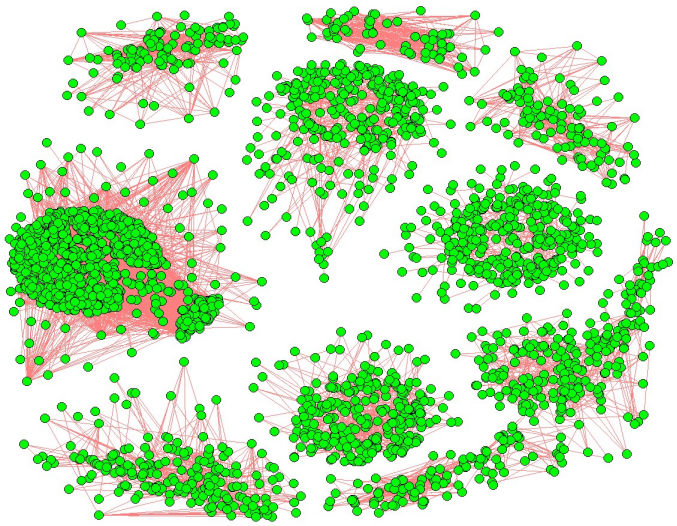

Figure 4.10: A snapshot of the components existing in GP dataset[UCINET 5].

	Number of			Size of	
Years	Exis author	New author	Components	Max compo.	Max deg.
1986	0	2	4	3	2
1987	2	7	5	3	2
1990	9	2	5	3	2
1991	11	1	11	4	3
1992	12	15	18	6	5
1993	27	17	44	7	5
1994	44	71	80	12	11
1995	115	119	121	21	15
1996	234	176	156	52	19
1997	410	159	197	119	29
1998	569	219	227	240	36
1999	788	222	253	314	43
2000	1010	192	280	434	47
2001	1202	223	317	522	58
2002	1425	302	342	623	62
2003	1727	232	376	788	62
2004	1959	311	414	922	79
2005	2270	302	439	1022	81
2006	2572	238	440	1022	81

Table 4.6: We juxtapose the new authors joining every year with the component statistics to get interesting information. For example, 238 new authors joining in 2006 do not join the largest component as the largest component is formed by 2005.

Years	Triangle-T1	Triangle-T2	Triangle-T3	Triangle-T4
1990	0	0	0	0
1991	0	0	0	0
1992	1	1	0	0
1993	3	5	0	1
1994	10	5	0	1
1995	33	8	1	1
1996	188	31	9	2
1997	260	50	22	2
1998	324	69	35	16
1999	452	115	56	25
2000	735	171	77	32
2001	981	216	91	53
2002	1615	300	128	82
2003	2247	362	177	142
2004	2433	431	209	161
2005	3356	497	278	212
2006	4081	566	319	274

Table 4.7: Triangle formation in the largest component of GP seems to be following the trend of the whole data set as seen in table: 4.3

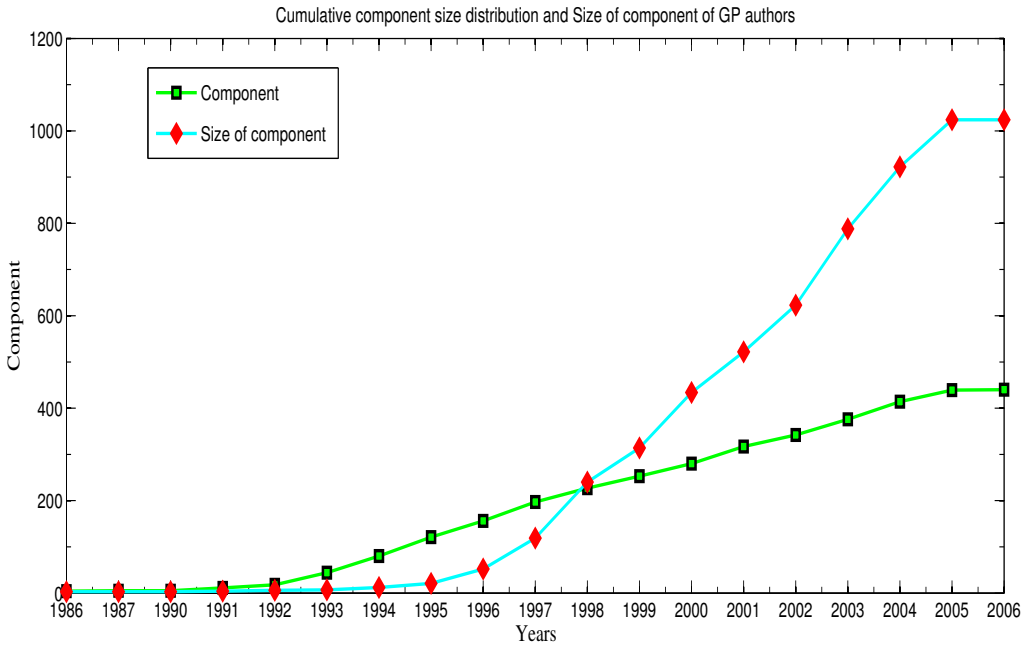

Figure 4.11: Number of componets growing every year and size of the largest of these components is shown. The new connections are seen to be attaching to the existing largest component. At the same time new small distinct components keep getting added to the network.

4.4.6 Rate of growth comparison with simulation results

The tertiary attachment model proposed in chapter 3 is simulated on the parameter set laid out in table: 3.1 to generate a network of size 1022. The growth of average degree of the actual GP component and that of the simulated networks are plotted in figure:4.12. For the simulation network, the year stamp is given by considering the number of edges present in GP network at that year. Similarly the growth of average clustering coefficient is plotted by simulating the model for the diffirent parameter sets for a network of size 1022. It can be clearly seen that the growth of GP component is modeled well by the sets 3 and 4 ie. by taking secondary contacts $\in$ U[0, 2] and tertiary contacts $\in$ U[0,2] or U[0, 1].

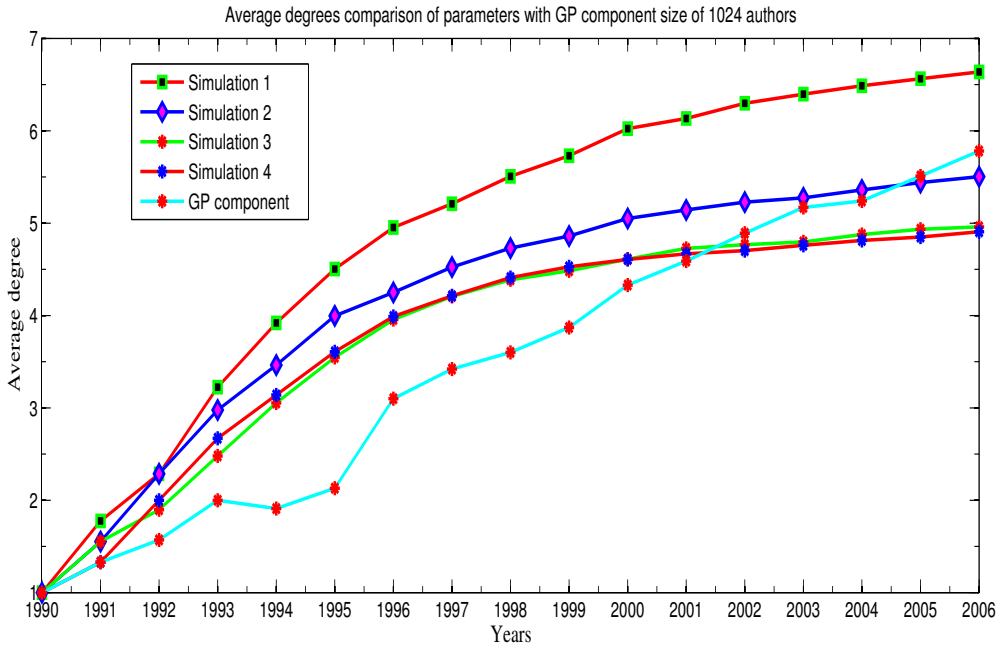

Figure 4.12: The growth of average degrees of GP authors component size of 1022 from 1990 to 2006, plotted along with simulation of TA model with parameter sets 1, 2, 3, 4 defined in table: 3.1.

The model is run with different parameters for primary, secondary and tertiary attachments with (parameter2) $m_r = 1$, $m_s = 3$ and $m_t = 2$ and (parameter3) $m_r = 1$, $m_s = 2$ and $m_t = 2$ are similar, see figures: 4.12 and 4.13. The average statistics of degree, clustering coefficent have been plotted for simulated and the actual real world network. We see that the parameter choice of parameter 3 is giving better result with $m_r = 1$, $m_s = 2$ and $m_t = 2$. Hence, we simulate the model with these parameters

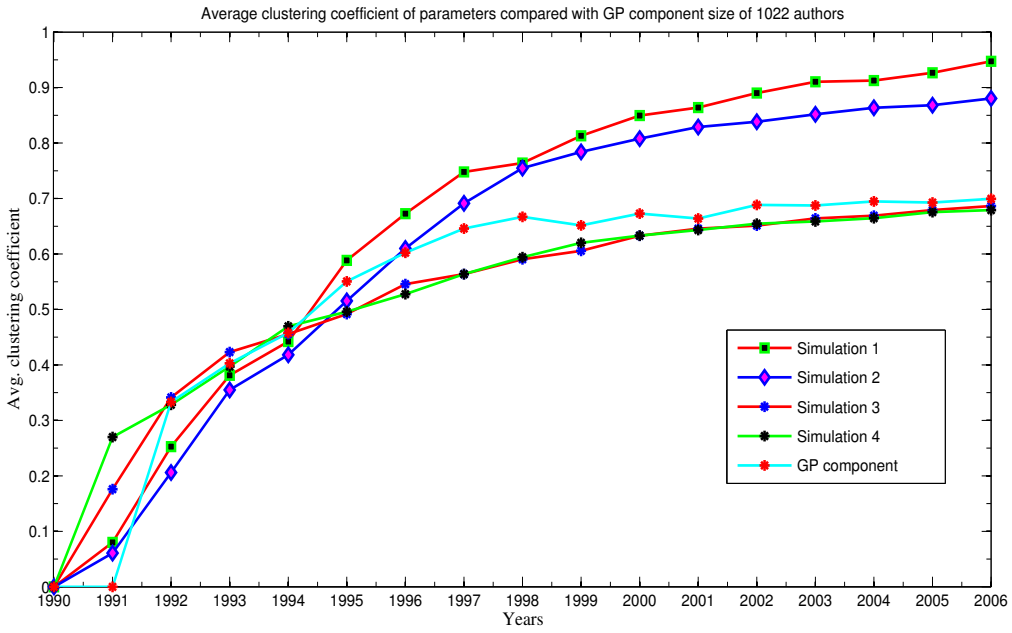

Figure 4.13: The growth of average clustering coefficient of GP authors component of size 1022 between 1990 to 2006. Simulation with parameter sets 3 and 4 seem to be fitting closely to the GP characteristics.

and compute the average statistics and present them in table: 4.8, this table shows that the average statistics obtained by the simulation of tertiary attachment model is very close that of GP component.

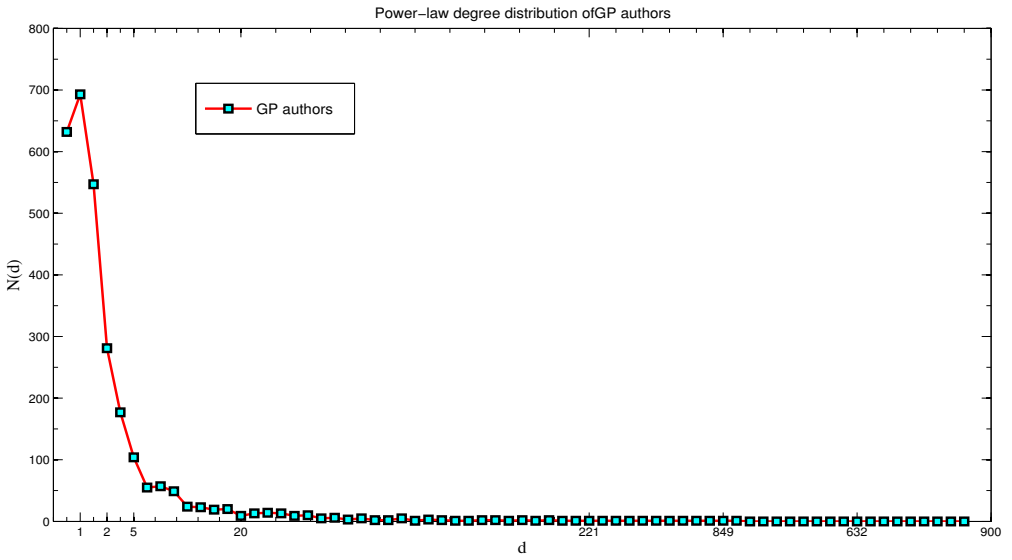

Figure 4.14: Above figure shows the degree distribution of GP author collaboration. It clearly indicates powerlaw behavior.

	GP component	Tertiary attachment model
Authors/users	1022	1022
Average degrees	5.72	5.78
Avg. clus. coefficient	0.674	0.699
Average path length	4.74	4.78
No. of components	1	1
Size of component	1022	1022
Maximum degree	81	73

Table 4.8: Comparison of average statistics of GP component dataset with those obtained from simulation.

4.5 Empirical analysis of Facebook dataset

For the analysis purpose we have considered three years of data from $facebook$ (FB) which is available with time stamp, between September 2006 to 2008, formation of edges given date wise. The data does not contain any authenticated user information. Tables: 4.9, 4.10 and 4.11 show the data with cumulative month wise record of new users introduced into the network, the number of new contacts and their interactions formed.

4.5.1 Facebook

Facebook was basically a printed or online directory given to students in American Universities for helping them get to know each other. FB website was first launched by Mark Zukerberg on 28th October 2003 in Harvard University. In 2006, FB launched the High school version of the website and on 26th September 2006 it was opened for everyone with age 13 or above with a valid e-mail address. By that time with the development of Web-2.0 technology, FB offered features like messaging, search for friends, friend groups, photo albums and profile photo history etc.

In 2007, more features were added to FB like mobile uploads, photo, video tagging and rss feeds, sms updates and a feature like virtual group was facilitated. By the end of 2007, FB took a new turn by launching social ads and marketing relevant advertisements and the feature of friends lists was launched. In 2008 FB became more active and incorporated the features of FB chat, friends suggestions and FB wall etc.

In 2010 FB concentrated mostly on speed and simplification of applications and in the enhancement and simplification of privacy options, also introduced new features like games and community pages. By the end of December 2012, FB became world's largest social network with 1.06 Billion Monthly Active Users (MAU) [http://en.wikipedia.org/wiki/Facebook]

4.5.2 Literature review on Facebook users

Lewis *et al.* (2008) provided a benchmark dataset for facebook. They have collected facebook dataset of students of a college in US and studied some of the useful properties of dataset and made it available for further reference and research.

Gjoka *et al.* (2008) studied the usage of third party applications in facebook. They have collected data pertaining to 180 days between August 2007 to February 2008 from different geographic regions and analyzed the usage of different applications in facebook as a function of number of days and number of users.

Gjoka *et al.* (2010) model using Markov chain random walk to obtain a uniformly distributed dataset and applied re-weighting Random Walk (RWRW) and Metropolis-Hastings Random Walk (MHRW) approximations and studied their efficiency in predicting the degree distribution of the users of Facebook.

A paper by Catanese *et al.* (2011) proposed a method to crawl the FB for data collection based on Breadth First Search (BFS) and analyzed degree distribution, centrality measures, scaling laws and distribution of friends etc. using the Stanford Network analysis library which provides general purpose network analysis functions.

Hu and Wang (2012) collected 800 days dataset of FB from May 2005 to August 2007 and studied the temporal behavior of request and acceptance of friendship in Facebook. They have also studied complimentary cumulative degree distribution and proposed a preferential attachment model which fits the observed behavior in the dataset.

Viswanath *et al.* (2009) studied the usage of facebook wall and its influence on the degree, clustering coefficient and average path length of FB dataset over a period

of time between January 2007 to January 2009. They have observed a considerable change after the introduction of new site design with convenient wall access starting on 20th July 2008.

Ferrara and Fiumara (2012) studied various model's applicability to real world social network datasets available online like Facebook, youtube, arXiv etc. and fitted the real world social network data with the model parameters. They studied power law of degree distribution, average path length etc. in order to understand whether the model is supporting small world characteristics, scale free property of degree distribution and community structure formation of the real world online social network data.

A recent work by Ferrara (2011); Ferrara *et al.* (2012) investigated the role of strong and weak ties in FB. In particular they have presented a quantitative analysis of "Strength of weak ties" Gramovelter's theory. They have used uniform sample datasets of 2009 FB with almost a million nodes and 55 to 75 millions of edges.

Traud *et al.* (2011) studied the social structure of FB network. Here they have analyzed friendship networks at one hundred American colleges and Universities at one point in time. They have studied assortivity nature and detected community structure and concluded how macroscopic and microscopic perspectives give complementary insights on social networks.

The above literature review papers are discussed based on friend of friend contacting a link and analysis done on average degrees, average clustering coefficient and average path length. With this background in mind we have applied our Tertiary attachment model to analyze the network structure of Facebook.

4.5.3 Facebook users dataset

The datasets considered fom our analysis are given in tables: 4.9, 4.10 and 4.11. In 2007 there were only 15 thousand records which constitute around 10 thousand users and 10 thousand edges, where as by 2008, there were around 60 thousand records which lead to around 25 thousand users and around 35 thousand edges.

Months	Input edges	Nodes count	Interactions
September-06	4818	3218	3780
October	12820	5816	9969
November	19699	7378	15320
December	26557	8620	20545

Table 4.9: Cumulative month wise Facebook users data along with edges and interactions.

Months	Input edges	Nodes count	Interactions
January-2007	8153	5278	6202
February	16311	7808	12393
March	26048	9773	19770
April	36043	11357	27298
May	47315	12834	35774
June	60452	14342	45578
July	75644	15898	56761
August	92889	17463	69561
September	110745	19295	82605
October	129746	21201	96222
November	148197	23105	109368
December	163383	24678	120176

Table 4.10: Cumulative month wise facebook users, input edges and interactions.

Months	Input edges	Nodes count	Interactions
January-2008	16361	10680	11580
February	32116	16135	22804
March	53740	20986	39238
April	87673	25198	39238
May	118002	28186	65353
June	148529	31243	126223
July	183995	34633	154889
August	222385	38309	186266
September	259034	41558	216062
October	301863	45039	251268
November	346356	48941	288141
December	406591	54161	323474

Table 4.11: Cumulative month wise facebook users, input edges and interactions.

This gives an idea of how with time the number of users doubled and edges tripled with number of records increasing by four times. ie; acceptance, deletion and reconsideration as friend has increased by 6 times which indicates the usability of FB by people. The dataset taken from Viswanath *et al.* (2009).

4.5.4 Formation of connections and triangles

We consider initial attachment as an edge formed by a new user to the existing network. Secondary attachment is identified as an edge formed between friend-of-friend ie. i and j formed a link on day t_1 and j and k made friends on $t_2 \geq t_1$, then an edge forming on $t_3 > t_2$, between i and k represents a friend-of-friend secondary attachment.

In Graph G5, user 594 connects with 429 and 14662 on 02/01/2007. Note the secondary attachment being made between users 14661 and 429 on 04/01/07, two days later. Similarly, Graph G6 in figure: 4.15 depicts tertiary connection when a new user 582 connects with 429 and 592 on 08/01/07 in the process 429 and 592 also making a friendship link.

Triangles T_1, T_2, T_3 and T_4 are as defined in section 4.4.3 with year replaced by day, ie. for eg. T_2 represents a triangle with two edges formed earlier and the third link forming on a later day.

We present two snapshots of facebook friendship subgraphs of different sizes in figures: 4.16 nad 4.17.

The definition of primary, secondary and tertiary connections is given in section: 3.4. Applying the definition of attachments, we extract the same from FB datasets and the results obtained are plotted in figures: 4.18 and 4.19.

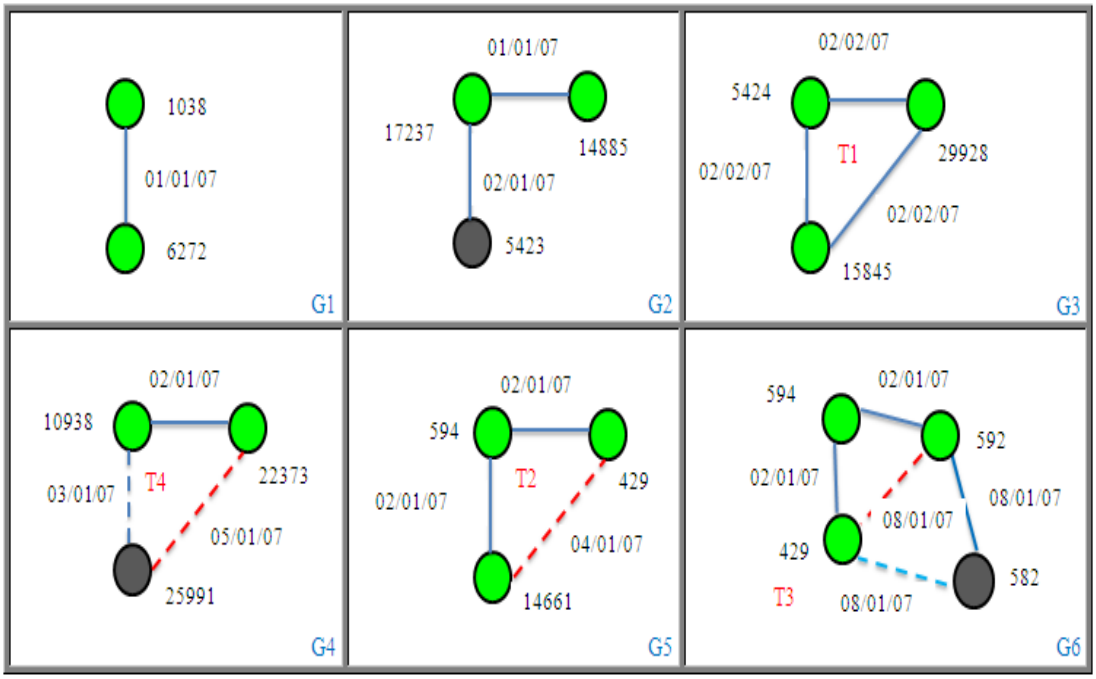

Figure 4.15: Different types of connections and triangles found in the FB dataset. The node id's are given along with the date on which the connections are made.

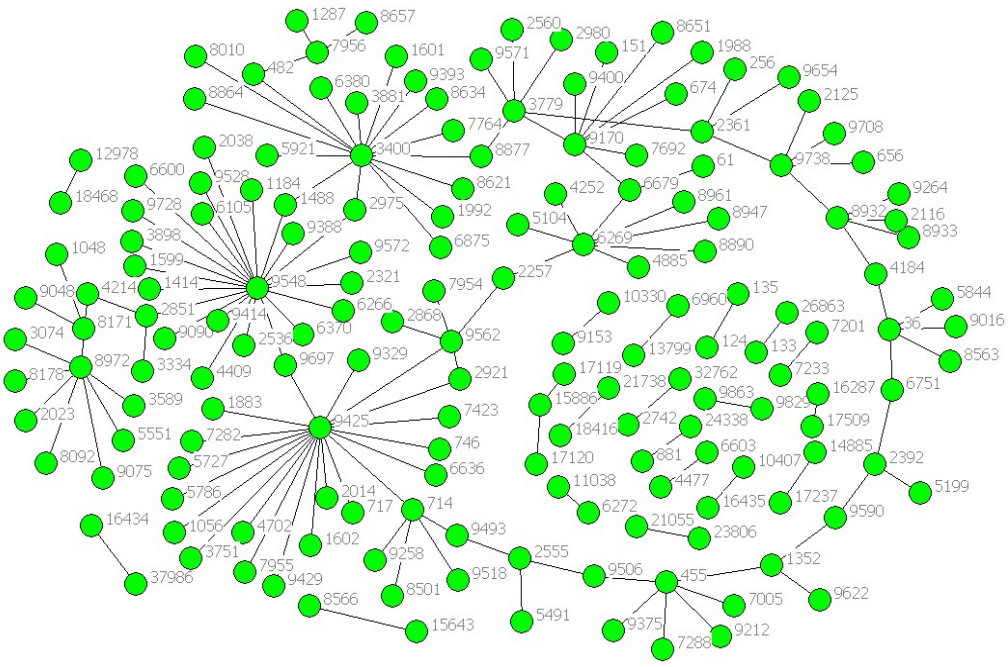

Figure 4.16: A snapshot of Facebook friendship network among 170 users in 2007.

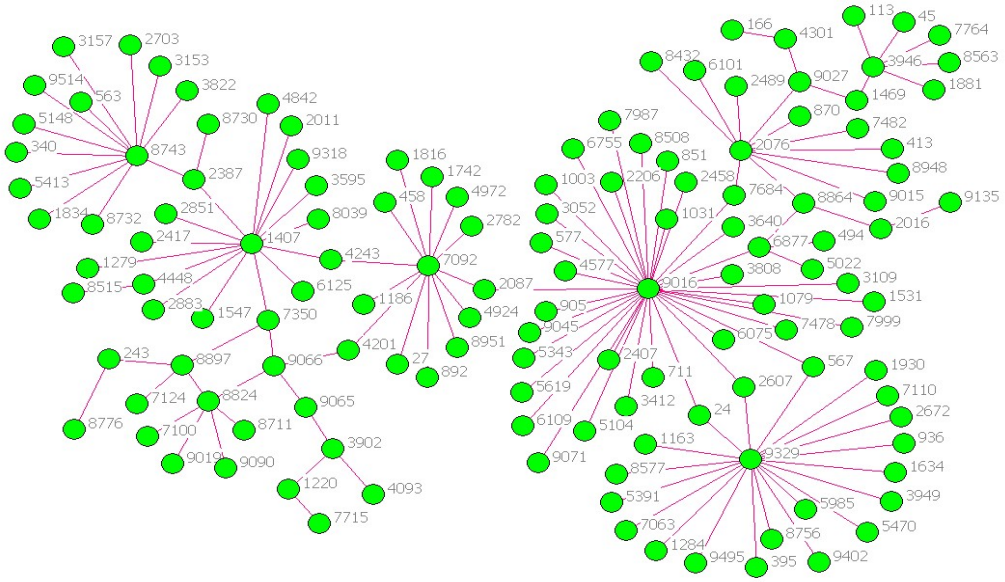

Figure 4.17: A snapshot of community structure in Facebook users among 135 users in 2007.

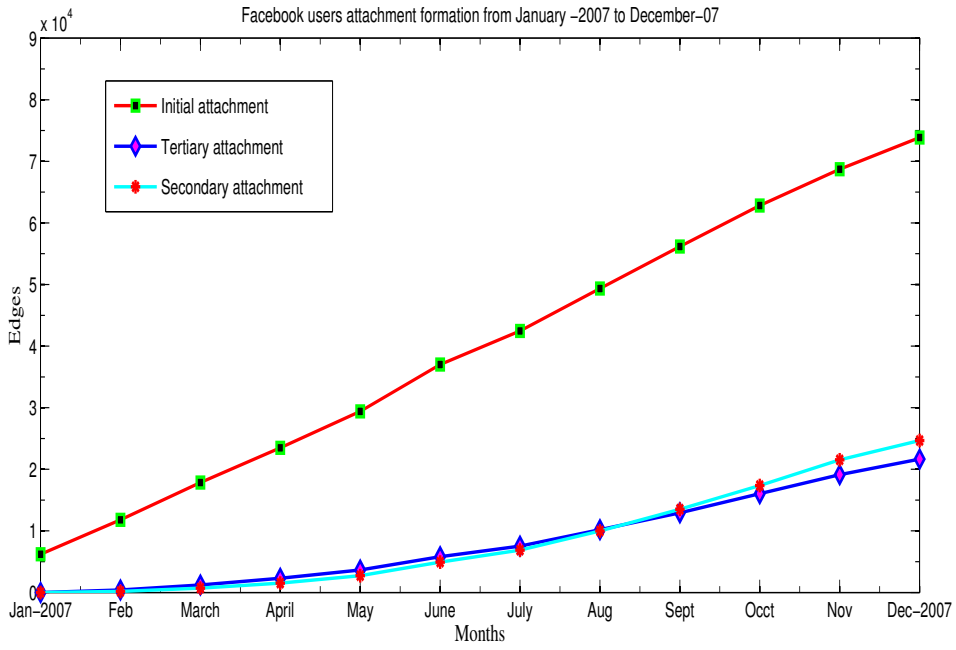

Figure 4.18: Plot of initial, secondary and tertiary connection count as a function of monthr for FB dataset in 2007.

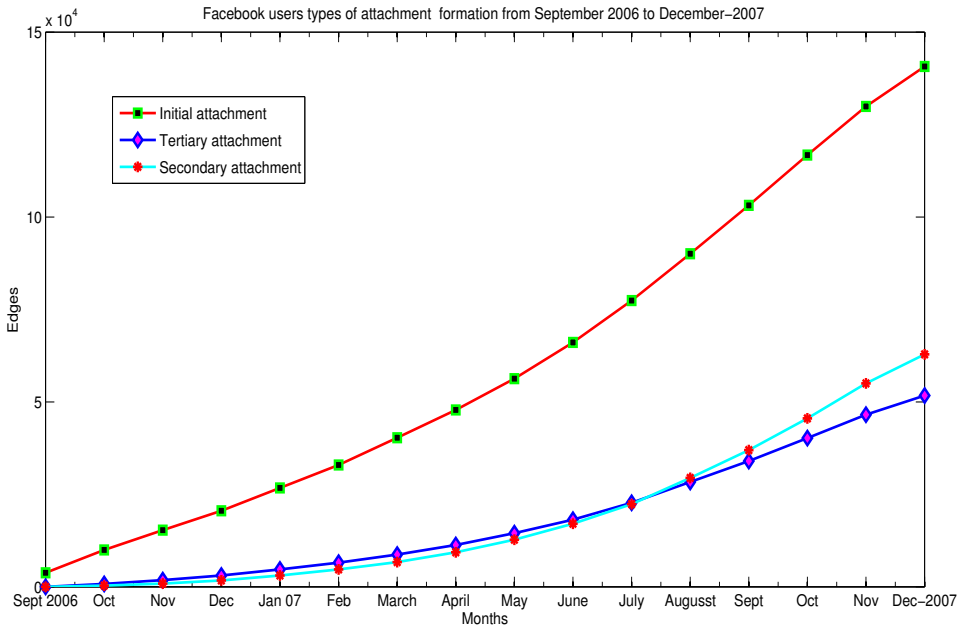

Figure 4.19: Plot of initial, secondary and tertiary connection count as a function of month for FB dataset from 2006 to 2007.

Facebook dataset is analyzed for primary, secondary and tertiary attachments. Different types attachments and triangle formations during September 2006 to December 2006 and September 2006 to December 2007 are presented in tables: 4.12 and 4.13. We do not present cumulative results for 2008 data as number of users and triangles is relatively huge.

	Edges			Triangles formation			
Months	Ini.attach.	Sec.attach.	Ter.attach.	T1	T2	T3	T4
September	3780	0	0	301	0	0	0
October	9126	479	364	838	547	130	0
November	13328	1152	840	1119	1145	400	225
December	17128	1957	1460	1384	1757	718	827

Table 4.12: Above table shows the average results of over 4 months from Sept-2006 to Dec-2006 for attachments and triangles.

In table 4.12 it can be seen that the count of T_1 and T_2 much higher than T_3 and T_4. Further T_2 which captures secondary attachments is much higher than primary. Tertiary attachment is also significant as proposed by the TA model. In figures: 4.20 and 4.21 the growth of different types of triangles is plotted.

	Types of Contacts			Triangles formation			
Month	Init.attac.	Sec.attac.	Ter.attac.	T1	T2	T3	T4
Sept-2006	3780	0	0	301	0	0	0
October	9126	364	479	838	547	130	0
November	13328	840	1152	1119	1145	400	225
December	17128	1460	1957	1384	1757	718	827
Jan - 2007	21371	2423	2953	1715	2451	1209	1705
February	25456	3414	4079	2007	3179	1693	3020
March	30104	4766	5471	2479	4093	2299	4587
April	34476	6409	6994	2835	5037	3096	6671
May	39146	8390	8797	3324	6116	3962	9517
June	44308	10919	10900	3951	7403	5151	12766
July	50030	14060	13350	4767	8917	6473	16900
August	55840	17933	16334	5589	10672	8191	22333
September	61829	22092	19247	6635	12434	10058	27808
October	67718	26576	22488	7440	14381	12103	34230
November	72997	31231	25688	8266	16203	14129	41321
December	77688	34745	28288	8683	17484	15617	47949

Table 4.13: Above table shows brief results of over period of one year four months results from September 2006 to Dec.2007. We can see number of contacts, types of attachments and formation of triangles is increasing.

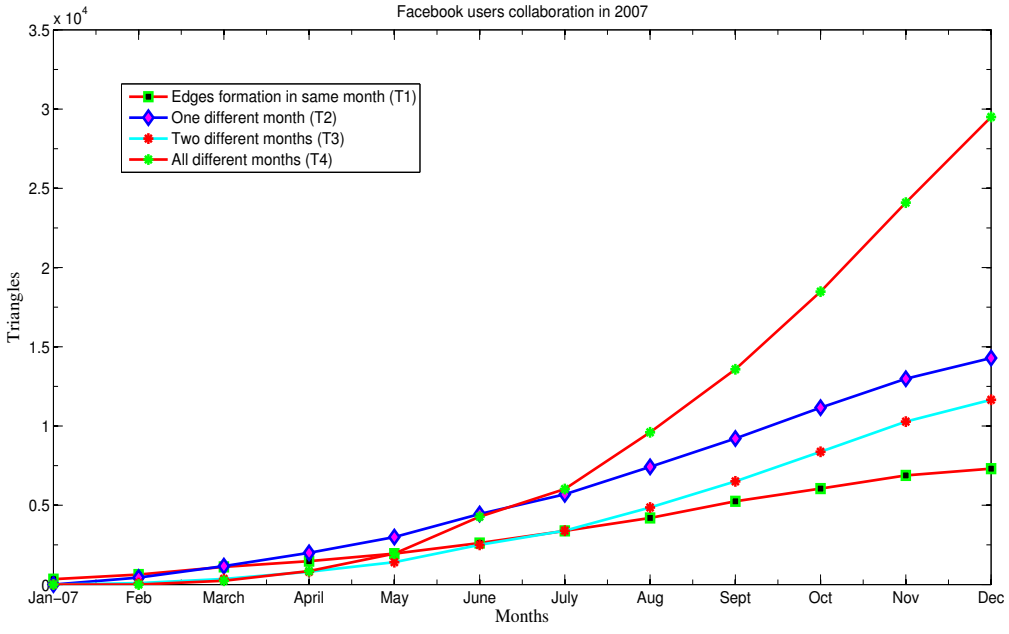

Figure 4.20: Plot of T_1, T_2, T_3 and T_4 over the year 2007.

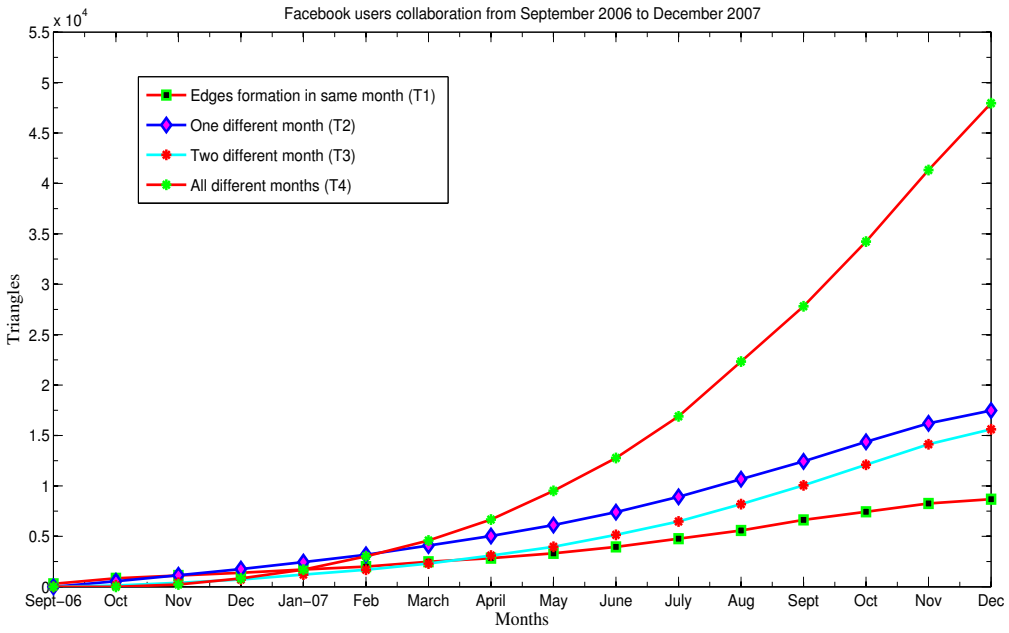

Figure 4.21: Plot of Triangle formation from September 2006 to December 2007.

4.6 Average statistics and component analysis

Average statistics of Facebook users shown in table: 4.14. A plot of number of components and growth of the size of largest component during the year 2007 and 2008 are presented in figures: 4.23 and 4.22 respectively. It can be seen that the number of components is decreasing indicating new connections merging the components. This is further validated in the next graph in figure: 4.22, which shows the size of the largest component increasing over the year.

The month wise cumulative whole year average degree, average clustering coefficient and average path length are presented in table: 4.15 respectively for the years 2007 and 2008. The statistics clearly show that the average degree increased considerably and clustering coefficient increasing during 2007 remained almost same during 2008. Where as the average path length decreased by half in the year 2007 and 2008. The invariance of clustering coefficient with an increment in average degree does not indicate a corresponding increment in number of triangles but decrement in average path length indicates the formation of more local communities hence more triangles.

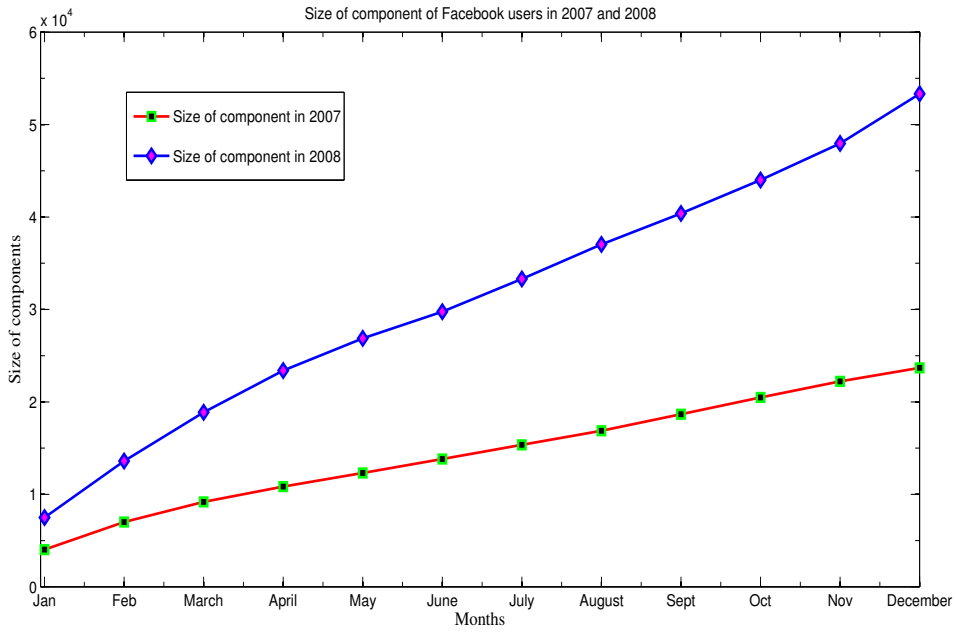

Figure 4.22: Size of component can be seen to be growing in both and the growth is observed to have a correlation 0.996.

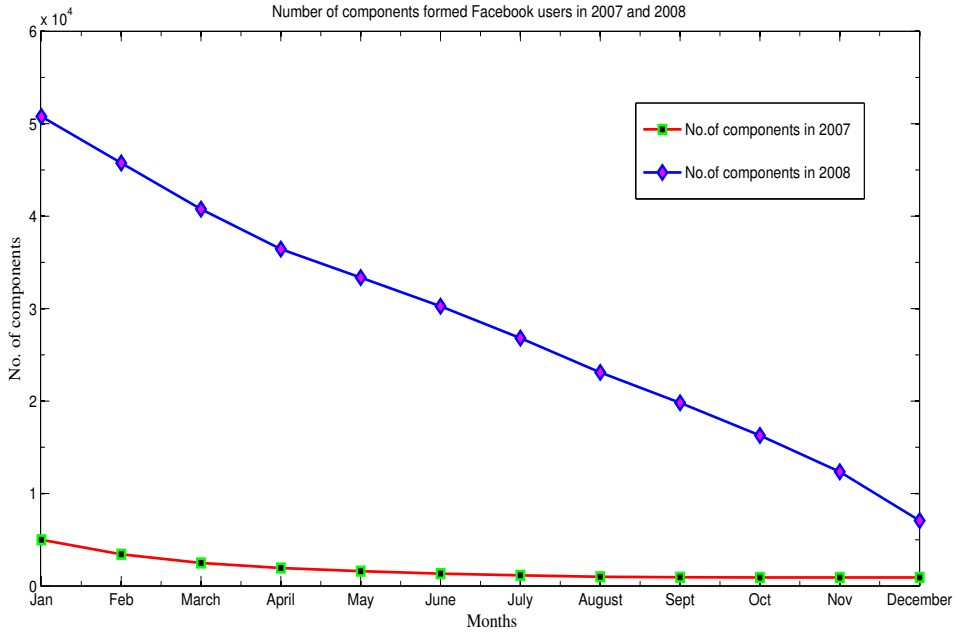

Figure 4.23: Plot of number of components during 2007 and 2008 of Facebook users.

The trend of different statistics for FB also follow an intuitive estimate. Average degree of the nodes keeps steady increasing (almost linearly), with clustering coefficient remaining very low where as path length is fairly stable in the range 5-7 which may indicate a six-degree separation.

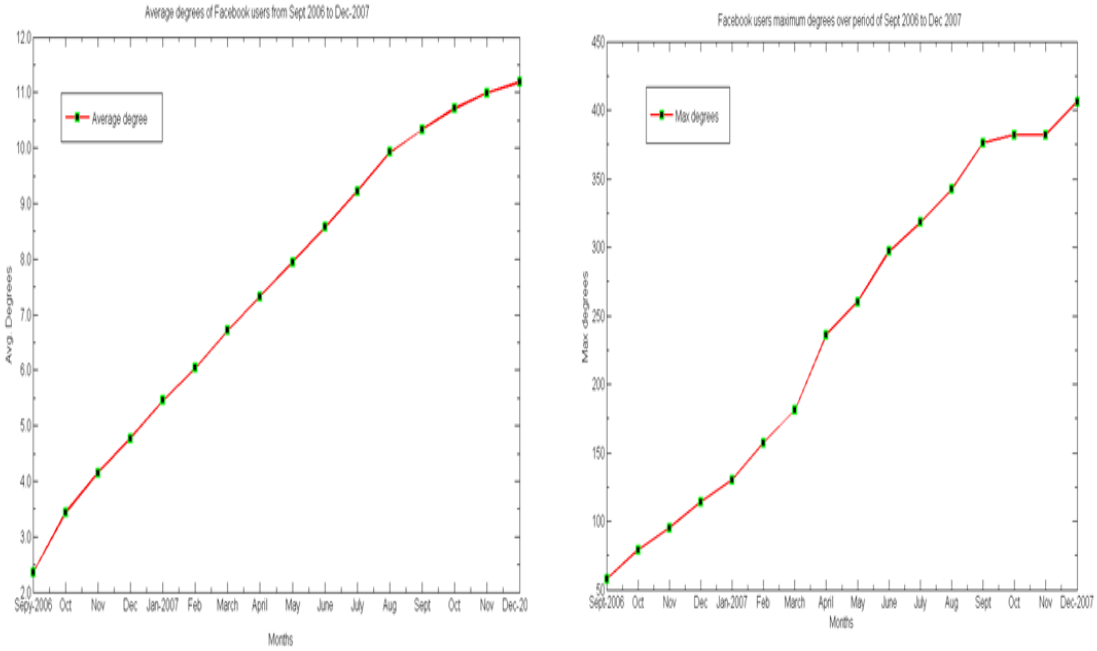

Figure 4.24: Months wise average degree and maximum degree formation of Facebook users dataset on 2006 to 2007.

	Averages from Sept-06 to Dec-07			Number of	
Month	Degree	Clust.Coeff.	avg.PathLen.	Components	Max deg.
Sept-2006	2.35	0.036	9.57	350	58
October	3.43	0.062	8.41	246	79
November	4.15	0.071	7.62	219	95
December	4.77	0.077	7.11	203	114
Jan - 2007	5.44	0.083	6.62	172	130
February	6.02	0.088	6.36	180	157
March	6.68	0.093	6.15	169	181
April	7.27	0.097	5.99	167	236
May	7.88	0.101	5.87	182	260
June	8.50	0.106	5.73	188	297
July	9.13	0.110	5.62	198	318
August	9.82	0.113	5.51	207	342
September	10.22	0.118	5.47	254	376
October	10.58	0.120	5.43	299	382
November	10.86	0.122	5.41	361	382
December-07	11.04	0.123	5.39	406	406

Table 4.14: Above table shows brief results of over period of one year four months results from September 2006 to Dec-2007, we can see average degree and averages clustering coefficient is increasing.

An analysis of cumulative whole year wise study of Facebook average results are presented in table:4.16. This shows that the average degree is increasing consider-

	Averages in 2007			Averages in 2008			
Months	Deg.	Clu.Coeff	Path Len.	Months	Deg.	Clu.Coeff	Path Len.
Jan-07	2.35	0.030	12.12	2Jan-08	3.16	0.027	13.89
Feb-07	3.17	0.044	9.02	Feb-08	3.98	0.030	9.56
Mar-07	4.05	0.059	7.81	Mar-08	5.21	0.040	8.02
Apr-07	4.80	0.068	7.24	Apr-08	6.95	0.056	6.96
May-07	5.56	0.077	6.83	May-08	8.37	0.065	6.48
Jun-07	6.49	0.085	6.45	Jun-08	9.50	0.070	6.16
July-07	7.12	0.091	6.19	July-08	10.62	0.072	5.91
Aug-07	7.93	0.095	5.96	Aug-08	11.61	0.073	5.73
Sept-07	8.48	0.102	5.83	Sept-08	12.46	0.075	5.61
Oct-07	8.98	0.106	5.73	Oct-08	13.40	0.077	5.49
Nov-07	9.36	0.109	6.67	Nov-08	14.15	0.077	5.39
Dec-07	9.63	0.111	6.62	Dec-08	15.01	0.077	5.33

Table 4.15: Comparison of the results of Facebook users two years on average degrees, average clustering coefficient and average path length of Facebook users.

	2006 (4 months)	2007	2006-07	2008	2006-08
Facebook users	8620	24968	25487	54161	56954
Average degrees	4.77	9.63	11.04	15.01	20.94
Average clust.coeff.	0.077	0.111	0.123	0.077	0.102
Average path length	7.11	6.62	5.39	5.33	4.93
No. of components	203	914	1081	7070	9521
Size of component	8180	23681	23571	53324	52414
Maximum degree	114	382	406	682	682
Density	0.0035	0.00286	0.00276	0.00138	0.00128
Assortative	0.0934	0.126	0.0684	0.0890	0.1212

Table 4.16: Cumulative analysis of Facebook users benchmark dataset from 2006 to 2008.

ably, is almost constant and low. Average path length decreases. A comparison in the trends of growth between years 2007 and 2008 of parameters like degree , clustering coefficient, average path length are plotted in figures: 4.25, 4.26 and 4.27. The degree distributions plotted in figures: 4.28 and 4.29 clearly exhibts powerlaw.

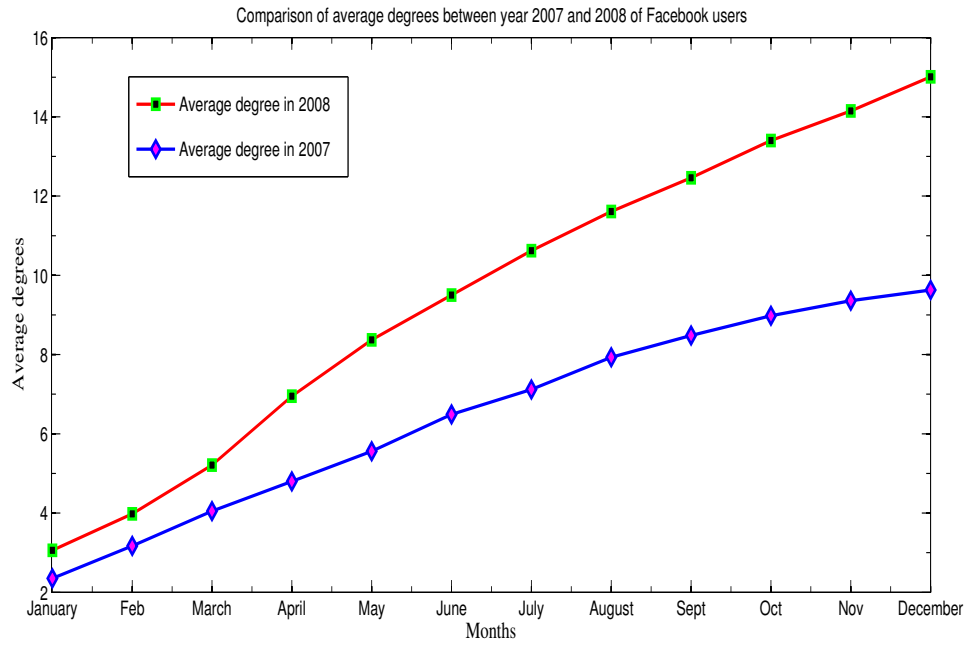

Figure 4.25: Year wise average degree formation of Facebook users dataset from 2007 and 2008. Mean value of average degree stands at 6.52 for the year 2007 and at 8.95 for the year 2008 with a difference of 1.52 and 0.998 correlation.

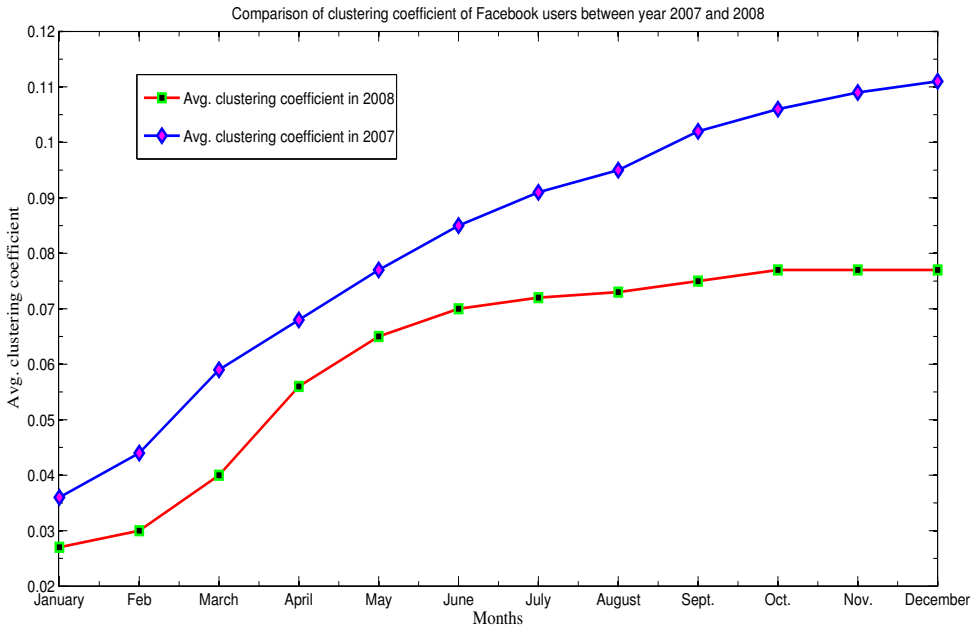

Figure 4.26: A low average clustering coefficient observed in Facebook users dataset during 2007 and 2008.

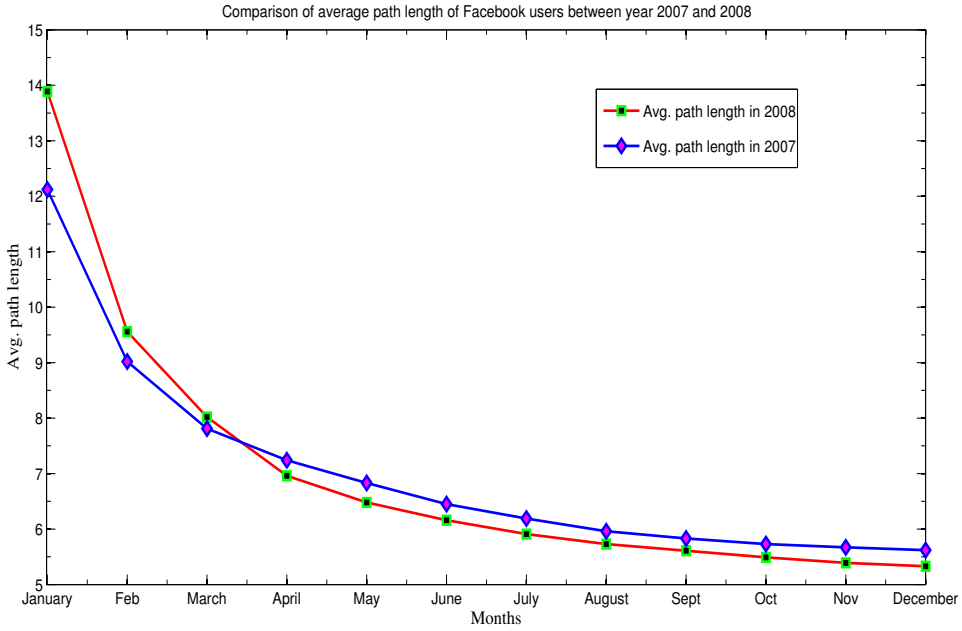

Figure 4.27: Average path length of Facebook users dataset show a similar profile in 2007 and 2008 with a correlation 0.99.

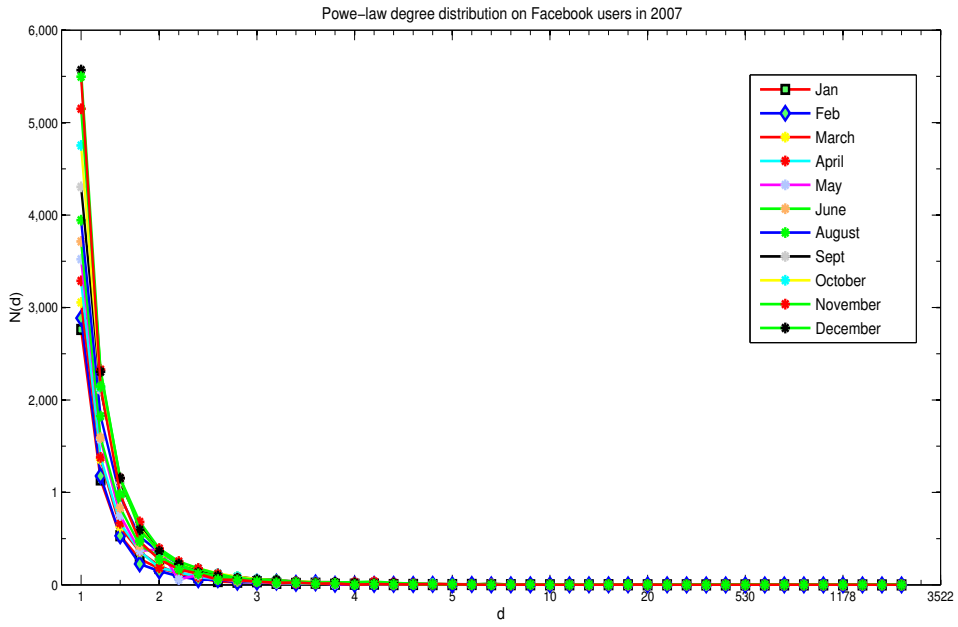

Figure 4.28: Plot of power-law degree distribution on one year of Facebook users in 2007.

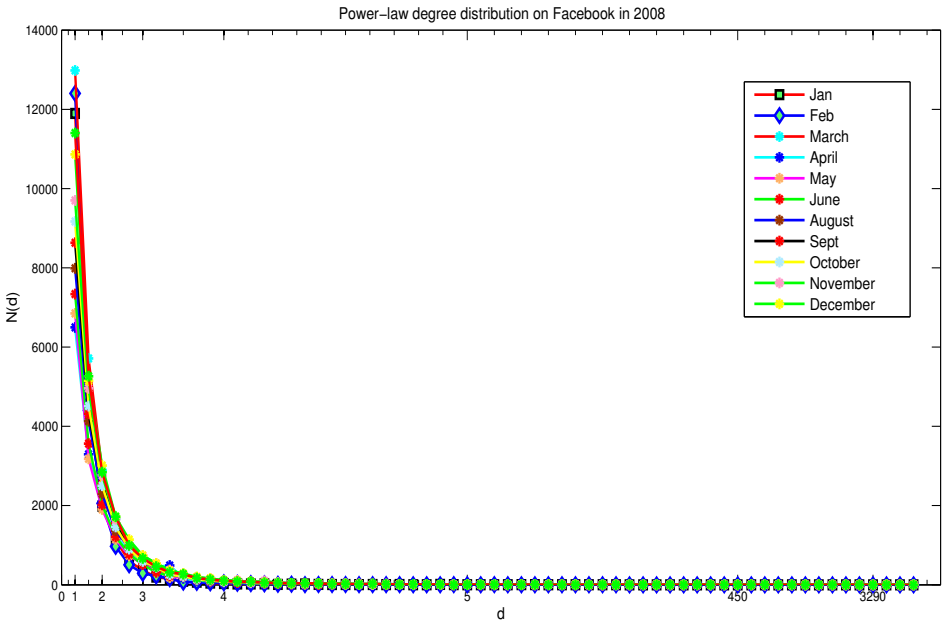

Figure 4.29: Plot of power-law degree distribution on one year of Facebook users in 2008.

4.7 Snapshot of a few other datasets

We give below a summary of average statistics of average degree, clustering coefficient etc. for a few other datasets like Bibsonomy, Netscience, Physica, DBLP along with GP is tabulated in table: 4.17. Eventhough most of these data sets are collaboration networks, their average statistics is quite different. Without the analysis of secondary of tertiary attachments, we still try to analyze for the strength of mutual collaborations in each of these networks.

It can be seen that the number of components in GP dataset is more but clustering coefficient is high and average path length is low. This shows the existence of mutual collaborations and a strong community structure in GP dataset, where as in Bibsonomy dataset the situation is quite opposite. The number of components is very less but the average path length is high and average clustering coefficient is very low. Also in Bibsonomy dataset, average degree is higher. This shows that the Bibsonomy dataset has more individual collaborations than mutual collaborations. ie; formation of triangles is low, but average degree is high. that is, there will be more

number of active centers in Bibsonomy dataset as compared to GP dataset. In Physica dataset, average degree is observed to be very high and number of components is relatively less. This indicates that in Physica community, there are more number of mutual collaborations when compared to any other dataset. Nect.Sci. and DBLP datasets have almost similar behavior as GP dataset, being all collaboration is carry out. Bibsonomy dataset is a user-tag bipartite graph with edge representing tag assigned by a user. This dataset during 1995 to 2005 is available. Bibsonomy is a web service in which the user can put their bookmarks. The node of 8823 represents a user with so many tags and other users be linked to some of these tags may belong to a different component. Hence we observe that this dataset possessesthe average path length between 1 and 2 and also low assortativity.

The neighborhood of the node a 7706 from Bibsonomy having highest degree of 8823 has been analyzed using 'R' software. Size of 1-hop neighborhood of 7708 is equal to 7500 and 2-hop neighborhood is 7000. Clearly this dataset consists of 7706 as a centre with almost all the nodes connected to it by first hop or second hop. Hence the average path length works out to be between 1 and 2.

Netscience dataset; Netscience is a highly focused collaboration network with scientists working in a common area called network science. So as expected, the statistics show a high number of 394 components which may represent seperate research groups in this area with high degree of collaborations within the group which is shown by low average path length and high clustering coefficient. Most likely also higher value of assortativity shows the inclination of a new researcher to join an established research group in this area. We verified that $Newman$ who is an initiator of this area of research is precisely the node of degree 34.

Physica and DBLP are similar, since the scientists from diverse area and as DBLP is a common repository not restricted one research area. Figure: 4.14 shows the degree distribution og GP author collaboration and that of Bibsonomy dataset. They clearly indicate powerlaw behaviour.

In this chapter we have presented a systematic analysis of various academic collaboration network datasets like GP, Bibsonomy, Net.Sci., Physica and DBLP. It has

	GP dataset	Bibsonomy	Physica	Netsci.	DBLP
Authors	2809	17320	1406	1589	19837
Avg. degrees	4.16	6.96	35.00	3.44	14.71
Avg clus coeff.	0.664	0.0068	0.644	0.695	0.688
Avg. path length	0.633	1.257	3.496	1.985	6.049
No. of compo.	439	106	22	394	4440
Size of compo.	1022	16823	1314	441	3607
Max. degree	81	8823	230	34	90
Density	0.00101	0.000202	0.0303	0.00108	0.00015
Assortative	0.225	-0.2038	0.3050	0.463	0.5141

Table 4.17: An empirical analysis of five benchmark datasets.

been observed that GP, Net.Sci., DBLP have almost similar behavior where as Physica exhibits more mutual collaborations and Bibsonomy dataset stands on the other extreme with very less mutual collaborations.

4.8 Conclusion

In this chapter, we investigate the applicability of the tertiary attachment model thoroughly in the context of two different types of social networks: a) collaboration networks and b) friendship networks. Genetic Programming(GP) data set and Facebook(FB) data set for which time stamps are available are chosen for this study.

We find that the total of secondary and tertiary attachments amount to nearly 40% of the total edges found. Analysis regarding the different kind of triangles which roughly correspond to primary, secondary and tertiary connections should be further studied from the point of view of predictive capability of the model. Using these ideas, if a part of the network is known, then with appropriate probabilistic analysis, the model can predict secondary or tertiary attachments between two given nodes.

The chronological evolution of FB is presented in the beginning of the chapter. For contrasting, we juxtapose the average values for different measures in the table: 4.18 for GP and FB.

It is interesting to note that

- An academic network like GP takes 20 years to grow to have 2809 authors whereas

	GP authors (1986-2006)	FB users (Sept-06 to Dec-07)
Authors / users	2809	25487
Average deg.	4.16	11.04
Average clust.coeff.	0.664	0.123
Average path len.	0.633	5.33
Components	439	1081
Size of component	1022	23571
Density	0.225	0.0276

Table 4.18: Comparison of the GP authors and FB users.

within 1 year 4 months, FB gains above 25,000 users.

- GP shows lower average node degree with a high clustering coefficient of 0.664, whereas FB exhibits an average degree of 11 with a very low clustering coefficient of 0.12.
- Interestingly, the friendship network does show a near 6-degree separation in the average path length computation whereas GP shows < 1 average path length which indicates denser graph with a possible community structure.
- Further, the largest component of GP constitutes 36 % of the whole, hence the other components also must be reasonably large. On the other hand, FB has a large component whose size is almost 92 % of the whole indicating really small other components.
- Finally, a snapshot of FB and GP are shown in figure: 4.30 capturing most of these interesting features.

Further, we show that the rate equations derived for the TA model, simulated with different parameter distributions, match very closely with the rate of change of degree and clustering coefficient in GP.

We study the interesting problems of community discovery and influence maximization in social networks in the next chapter. We can see in figures:4.30, that Facebook and GP data sets have significant community formation.

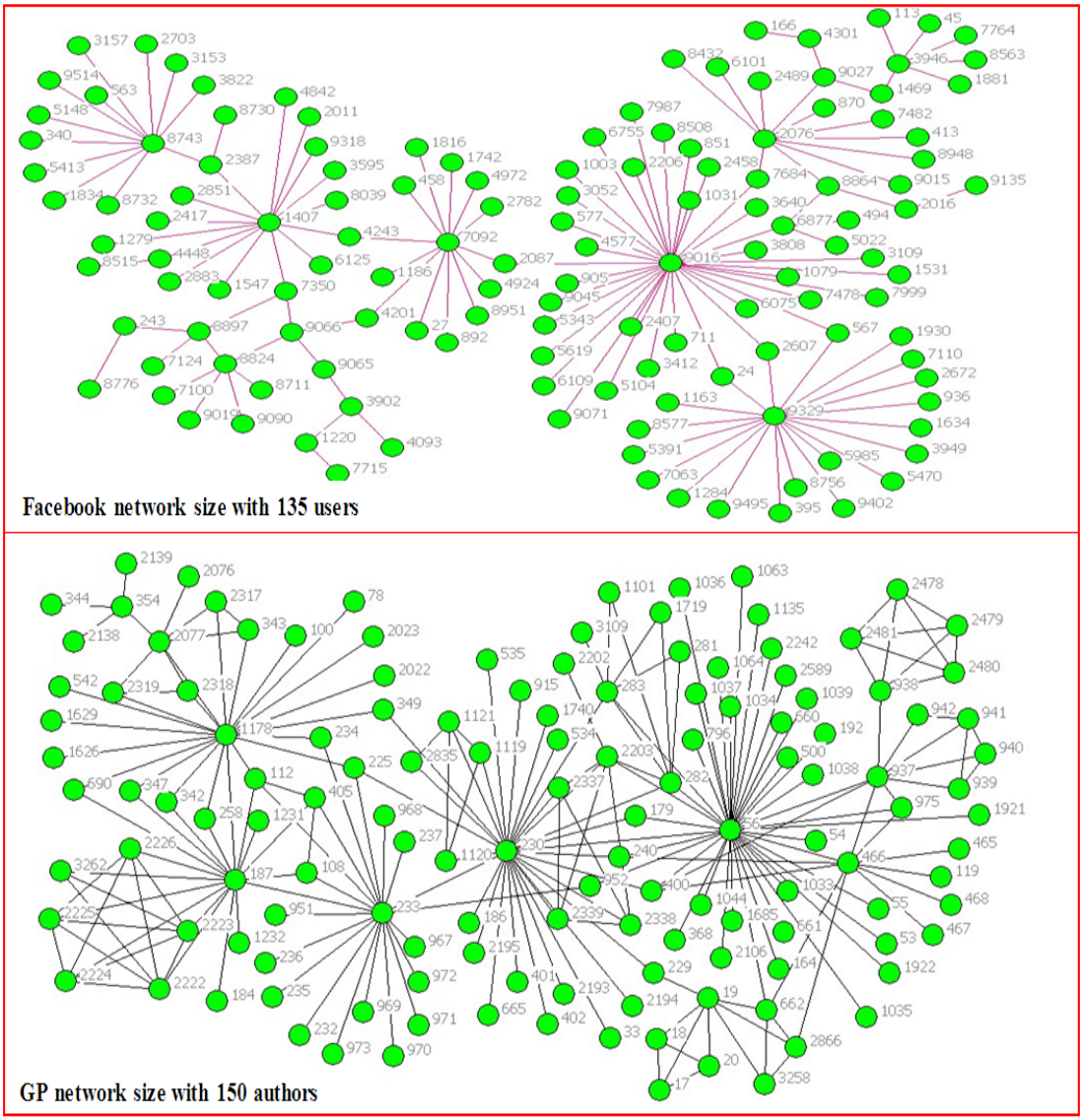

Figure 4.30: Snapshots of Facebook network with 135 users and GP network of 150 authors. Note the distinct triangles formations in GP whith very sparse star-like components in FB with almost no triangles.

CHAPTER 5

COMMUNITY DISCOVERY IN SOCIAL NETWORKS

5.1 Introduction

Community structures are quite common in real networks. Social networks often include community groups based on common location, interests, occupation, etc. Communities are groups within a network where they are connected more densely among themselves and sparsely with vertices outside the group. The community detection problem becomes challenging if there are more overlapping links between communities. Community discovery is a major research area in social networks.

There are several approaches to identify the community structure where each method relies on one of the distinguishing features of the network. Community discovering algorithms can be broadly classified into two categories namely, Agglomerative algorithms where similar nodes are clustered together starting from a null graph; and Divisive algorithms where edges joining dissimilar nodes are removed iteratively.

5.1.1 Agglomerative Method

Agglomerative methods use ideas of node-independent paths or edge-independent paths for a similarity measure. Clauset (2005) propose a local community structure which is called local modularity, which works as a fast agglomerative algorithm that maximizes the local modularity in a greedy fashion.

Newman (2006) proposed community structure based on modularity, initially it divides into two communities and further into more communities. Raghavan *et al.* (2007) proposed community detection using label propagation algorithm.

Chen *et al.* (2010) proposed a new local algorithm based on node strength to detect the overlapping community structures. The main strategy is to find an initial community from a node with maximal node strength and to expand the partial community from the initial one by adding nodes that are tight with the community.

Li *et al.* (2012) proposed an algorithm based on neighborhood overlap for community identification in complex networks. Fu *et al.* (2012) proposed a scalable community discovery method based on a threshold random walk strategy.

5.1.2 Divisive method

One of the primary divisive methods is the simpler traditional method based on graph partitioning algorithm. The graph is split into disjoint sets of roughly equal size using some modularity measure.

Girvan and Newman (2002) (GN) propose a simple way to identify communities in a graph by detecting edges that connect vertices of different communities; removing these will make the clusters get disconnected from each other. This is the philosophy of divisive algorithms and GN is the most representative method of divisive methods. It is based on the edge betweenness score that measures the fraction of all shortest paths passing through given link (Girvan and Newman, 2002; Du *et al.*, 2008). By removing links with high betweenness, we can progressively split the whole network into disconnected components, until the network is decomposed into communities consisting of singleton nodes. GN algorithm is considered an efficient algorithm for detecting communities that gives exact communities without disturbing internal edges within the community.

Brandes (2001) proposed algorithm for calculating faster edge betweenness score thus improving the complexity of GN algorithm.

Spectral properties of graph matrices are frequently used to find partitions. Traditional methods are in general unable to predict the number and size of the clusters, which instead must be fed into the procedure. Dongen (2000) proposed a cluster algorithm for graphs called 'Markov Cluster Algorithm' (MCL), an elegant method

based on the eigenvectors of the Laplacian matrix. The values of the eigenvector components are close for vertices in the same community, so one can use them as coordinates to represent vertices as points in a metric space. So, if one uses M eigenvectors, one can embed the vertices in an M-dimensional space. Communities appear as groups of points well separated from each other.

Here we have considered and mainly focused on GN community detection algorithm based on edge betweenness and we proposed an enhancement called Enhanced-GN(EGN) algorithm to make the GN algorithm more efficient.

5.1.3 Literature review on Betweenness

1. Freeman (1977) for the first time discusses point centrality and betweenness centrality calculation for nodes in a graph or network.

2. For community detection, Radicchi *et al.* (2004) proposed a self-contained version of the GN algorithm. They consider the edge-clustering coefficient, defined as the number of triangles to which a given edge belongs and thus detects the communities. This algorithm is not suitable for sparser graphs for detecting communities.

3. Gregory (2007) extend GN algorithm with specific method of deciding when and how to split vertices, named as new CONGA (Cluster-Overlap Newman Girvan Algorithm). This algorithm works by splitting vertices and computes split betweenness, have time complexity $O(m^3)$ in the worst case. CONGA algorithm is good but extremely slow.

4. Gregory (2008) describes a CONGA-Optimized algorithm which works based on local betweenness calculation. This calculation for edges is very expensive and CONGA algorithm is performed by a breadth-first search from every vertex. This algorithm is also expensive.

5. Szczepanski *et al.* (2012) proposed a community detection algorithm based on shapely value for calculating betweenness centrality, which is closely related to stress centrality. It has same complexity as the best known algorithm due to Brandes (2001) for detecting community based on edge betweenness.

6. One of the interesting algorithms proposed by Lee *et al.* (2012) is a Quick algorithm for Updating Betweenness centrality (QUBE). This algorithm updates betweenness centrality when a new edge joins the graph. This paper mainly focuses on when to update betweenness centrality in a graph and when not to update. They introduce a Minimum Union Cycle (MUC) using which edge betweenness is updated. This idea makes it much faster than Brandes (2001) algorithm for calculating betweenness.

Efficient implementations of GN algorithm are of time complexity $O(mn)$. We propose an enhancement to GN -(EGN) algorithm which is presented in this chapter.

5.2 Problem definition

A network is said to have community structure if the nodes of the network can be easily grouped into sets of nodes such that each set of nodes is densely connected internally and sparser connections exist between groups.

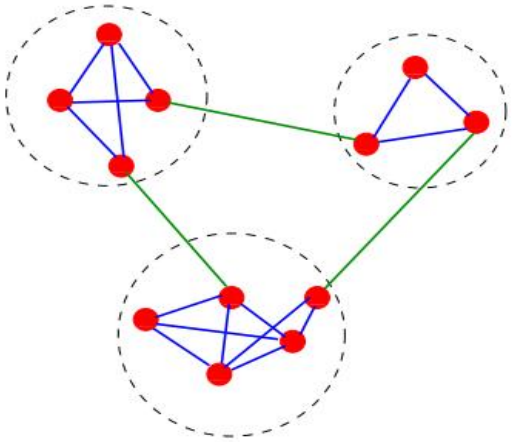

Figure 5.1: A snapshot of three communities within the circle, it has three betweenness edges between communities.

5.2.1 Betweeness centrality measure

The betweenness centrality measure plays a key role for community discovery. If one removes the bridge edges with highest betweenness score then the graph is divided into two communities. Let $\sigma_{u,v}(i,j)$ denote the number of shortest paths from u to v containing the edge (i,j) and $\sigma_{u,v}$ denotes the number of shortest paths from u to v in the undirected graph G. Then edge betweenness $C_B(i,j)$ is defined as

$$C_B(i,j) = \sum_{u,v \in E, u<v} \sigma_{u,v}(i,j)$$

Normalized betweeness centrality measure:

$$C'_{\mathrm{B}}(i,j) = \frac{2C_B(i,j)}{n(n-1)}$$

5.2.2 GN algorithm

GN algorithm is considered as one of the best algorithms for community discovery. Even though the algorithm runs in the worst case for $O(m\,n)$, where m is edges and n is number of nodes, it is supposed to detect communities exactly with very few internal edges disturbed. GN algorithm is described below.

1. Calculate the betweenness score for all edges in the network.
2. Remove the edge with highest betweenness.
3. Recalculate betweenness for all edges effected by the removal.
4. Repeat from step 2 until no edges remain.

5.3 Enhanced GN algorithm

Proposed enhanced algorithm works based on edge betweenness, with the idea that not one but many edges can be removed during a single iteration of GN algorithm.

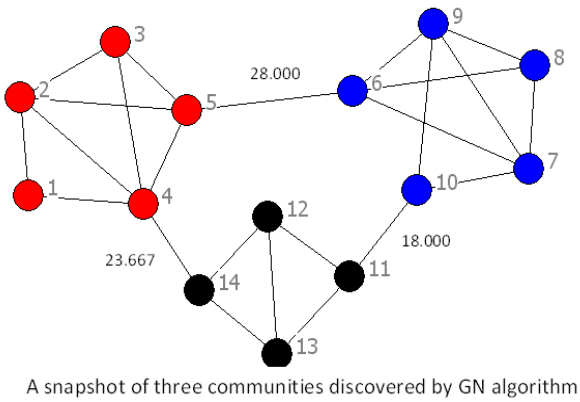

Figure 5.2: A snap shot of three communities detected as per GN, which takes three iterations to remove three bridge edges

5.3.1 Motivation

In the figure: 5.2 GN algorithm takes three iterations for detecting the three communities. Can we not remove the top three edges of edge betweenness score in the same iteration?

We propose in enhanced algorithm, to start with the calculation of the betweenness scores for all edges and remove the edge with highest score. And then among the next few edges with the highest scores, remove only those which are not part of a triangle.

Algorithm 2 Enhanced community discovery algorithm on a graph G. M is edge list of G sorted according to betweenness score in descending order, k is number of communities and t is threshold value.

```
1:  procedure (Betweeness Calculation )
2:      Calculate Betweenness for all edges of G
3:      Remove edge with highest betweenness from M
4:      for (i = 1; i ≤ t) do
5:          if (Edge (u_i, v_i) ∈ M have no common neighbor in G, then
6:              remove (u_i, v_i) from M and update G
7:          end if
8:      end for
9:      j = number of connected components of (G)
10:     if (j ≥ k) then
11:         exit
12:     else
13:         repeat step 2 to 10.
14:     end if
15: end procedure
```

This method will not effect any internal edges. Also this method is faster compared to GN algorithm.

Calculation of threshold 't' :

1. Compute betweenness scores for all edges e_i, $i = 1,2,\ldots,m$. Sort b_i to obtain $b_1, b_2 \ldots. b_m$.
2. Calculate the deviation of b_i from the second highest score, that is, $d_i = |b_2 - b_i|$, $i = 2,3,\ldots,m$
3. Normalize the deviation with maximum deviation $= \frac{d_i}{max d_i}$
4. $t = \#\{$ edges $e_i : d_i \leq 0.3 \}$.

5.3.2 Experimentation and results

We carried out analysis of community discovery for benchmark datasets of Zachary karate club, Santa Fe- collaboration network and Dolphin social network and compared the performance with GN algorithm which is taken to be the ground truth.

5.3.3 Datasets

Zachary karate club: This is the well-known Zachary karate club network. Each node represents a member of the club, and edges represent interactions. This is a classical social network dataset from the literature. The interactions among the members are supposed to reveal a political division among the members due to rivalry between two leaders of this club Girvan and Newman (2002).

Santa Fe collaboration network : Santa Fe institute collaboration network is extracted by Newman by considering the mutual collaborations between scientists in his institute Santa Fe Girvan and Newman (2002).

Dolphin network: An undirected social network of frequent associations between dolphins in a community. This bottlenose dolphin network is composed of 62 dolphins, edges represent interactions between dolphins and is composed of two communities of 41 and 21 dolphins each.

The basic statistics of the datasets are given in tables: 5.1 and 5.2, these dataset downloaded from url [http://www-personal.umich.edu/ mejn/netdata/].

Datasets	No. vertices	No. edges
Zachary karate Club	34	78
Santa Fe-Collaboration	118	221
Dolphin Social network	62	157

Table 5.1: The table gives benchmark data sets for community discovery

Datasets	Average degree	Avg. clus coeffic.	Avg. path length
Zachary Club	4.59	0.57	2.40
Santa Fe-Collabor.	3.75	0.58	5.10
Dolphin network	4.98	0.24	3.41

Table 5.2: A statistical summary of the three datasets of Zachary club, Santa Fe-collaboration and Dolphin social networks.

5.3.4 Implementation and results

Both GN algorithm and our proposed algorithm are run on the benchmark datasets and we compare the time taken in terms of number of iterations needed to discover communities as well as the correctness of the community composition. We consider GN as our ground truth which is presented in the table: 5.3.

	Zachary		Santa Fe			Dolphin		
Itere.	Edges	Score	Itere.	Edges	Score	Itere.	Edges	Score
1	32-1	71.39	1	78 - 64	3080.00	1	SN100- Bees	282.95
2	3-1	66.89	2	37 - 34	1394.00	2	SN9 - DN63	314.32
3	9-1	77.31	3	34 - 27	1392.00	3	Osca- Bees	360.09
4	34-14	82.03	4	41 - 37	296.00	4	SN100- SN89	531.57
5	34-20	123.23	5	64 - 41	280.50	5	PL - DN63	442.30
6	33-3	100.20	6	63 - 41	283.33	6	PL - Knit	861.00
7	31-2	143.62	7	65 - 41	248.33			
8	9-3	95.08	8	104 -78	201.08			
9	28-3	122.40	9	102 -78	170.20			
10	29-3	171.16	10	106 - 78	286.70			
11	10-3	288.00	11	103 - 78	221.36			
			12	106 -98	263.35			
			13	96 - 78	396.00			
			14	57 - 41	135.00			
			15	52 - 41	162.00			
			16	41 - 39	96.00			
			17	41 - 40	176.00			

Table 5.3: The edge with highest betweenness score is shown for every iteration of GN algorithm after updating the graph for the three datasets.

The first iteration of the Enhanced-GN algorithm on of Zachary karate club data set is presented in table: 5.4 and a summary of next few iterations of the algorithm are presented in table: 5.5

1^{st} Iteration	Top edges	Edge score	Deviation	Normalization value
1	**32–1**	**71.393**	**0.00**	**0.00**
1	7–1	43.833	0.00	0.00
1	6–1	43.833	0.00	0.00
1	3–1	43.639	0.194	0.004
1	9–1	41.648	2.185	0.055
1	**34–14**	**38.049**	**5.784**	**0.147**
1	**20–34**	**33.313**	**10.52**	**0.268**
1	12–1	33.000	10.83	0.276
1	27–34	30.457	13.37	0.341
1	5–1	29.333	14.50	0.369
1	11–1	29.333	14.50	0.369
1	13–1	26.100	17.73	0.452
1	25–32	23.594	20.23	0.516
1	28–3	23.109	20.72	0.528
1	22–1	22.510	21.32	0.543
1	18–1	22.510	21.32	0.543
1	26–32	22.500	21.33	0.543
1	23–34	19.489	24.34	0.620
1	21–34	19.489	24.344	0.620
1	19–34	19.489	24.344	0.620
1	16–34	19.489	24.344	0.620
1	15–34	19.489	24.344	0.620
1	24–34	18.328	25.505	0.650
1	31–2	18.110	25.723	0.655
1	2–31	18.110	25.723	0.655
1	10–3	17.281	26.552	0.677
1	30–34	16.722	27.111	0.691
1	17–6	16.500	27.333	0.696
1	8–3	14.145	29.688	0.756
1	29–34	13.781	30.052	0.766
1	4–3	12.583	31.25	0.796
1	33–34	4.6140	39.219	1.000

Table 5.4: Implementation of the Enhanced-GN algorithm on Zachary karate club. In the first iteration, the edges $32-1$, $34-14$ and $34-20$ are all removed as their deviation is < 0.3 other top edges 7-1 are part of a triangle and hence not removed.

Iteration	Top edges	Edge betweeenness score	Normalized- deviation
2	1 –9	91.28	0.00
2	33–3	76.87	0.0662
3	3 –1	126.56	0.00
3	31–2	87.46	0.00
3	28–3	74.24	0.1512
3	3 –9	121.00	0.00
4	3–29	171.16	0.00
5	34–10	117.83	0.00

Table 5.5: EGN algorithm needs only 5 iterations where as GN takes 11 iterations for Zachary karate club.

Iteration	Nodes	Edge betweeenness score	Normalized- deviation
1	78–64	3080.00	0.00
1	27–34	2552.00	0.15131
2	37–41	296.00	0.00
3	64–41	248.33	0.00
4	65–41	248.33	0.00
4	41–63	248.33	0.00
5	78–104	226.11	0.00
6	102–78	206.91	0.00
7	78–106	302.86	0.00
8	78–103	235.80	0.00
9	98–106	245.50	0.00
9	96–78	168.50	0.00
10	41–57	135.00	0.00
11	41–52	162.00	0.00
12	34–37	100.00	0.00
13	39–41	96.00	0.00
13	40–41	80.00	0.13043

Table 5.6: Proposed algorithm needs 13 iterations where as GN needs 17 iteration to detect the community.

Itere.	Edges removed	Edge scores	Deviation	Normalized deviation
1	SN100-Beescratch	293.4097	0.00	0.00
1	SN9-DN63	224.77	0.00	0.00
1	SN100-SN9	181.017	43.75	0.1946
2	Oscar-Beescratch	671.624	0.00	0.00
3	PL-DN63	526.168	0.00	0.00
3	Pl-Knit	375.833	46.21	0.109

Table 5.7: Proposed algorithm saves 50% of time, since it needs 3 iterations where as GN run for 6 iterations on Dolphin network.

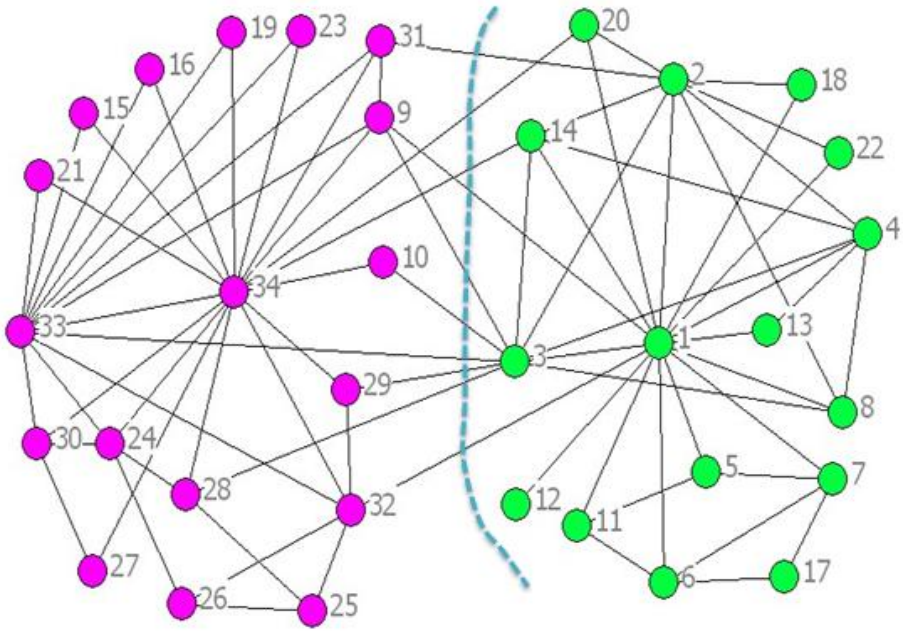

Figure 5.3: The friendship network of Zachary karate club of 34 members is divided in two communities and node number 10 comes into second community without disturbing internal edges.

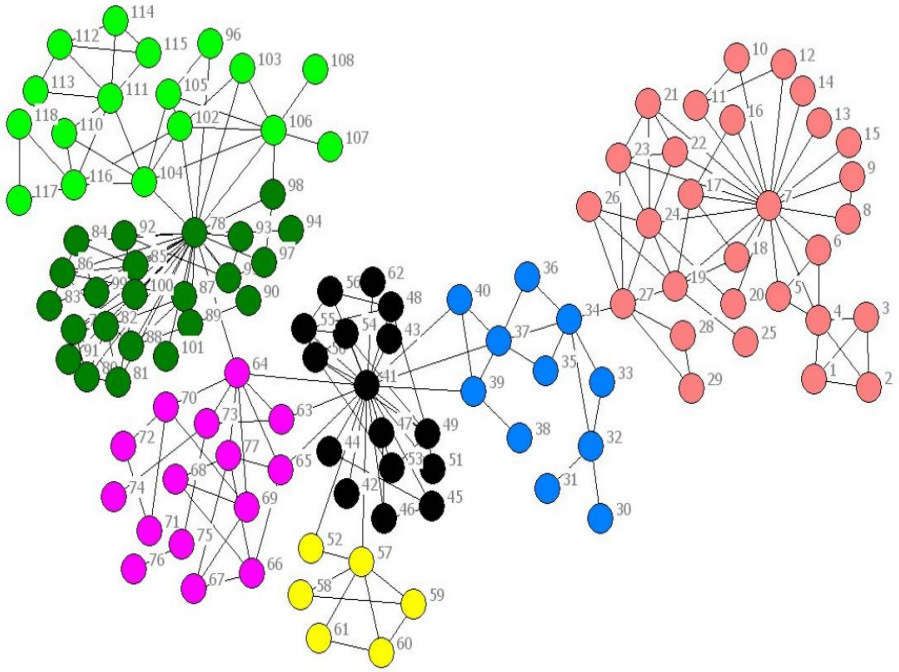

Figure 5.4: Santa Fe collaboration network having 7 communities depicted in different colours.

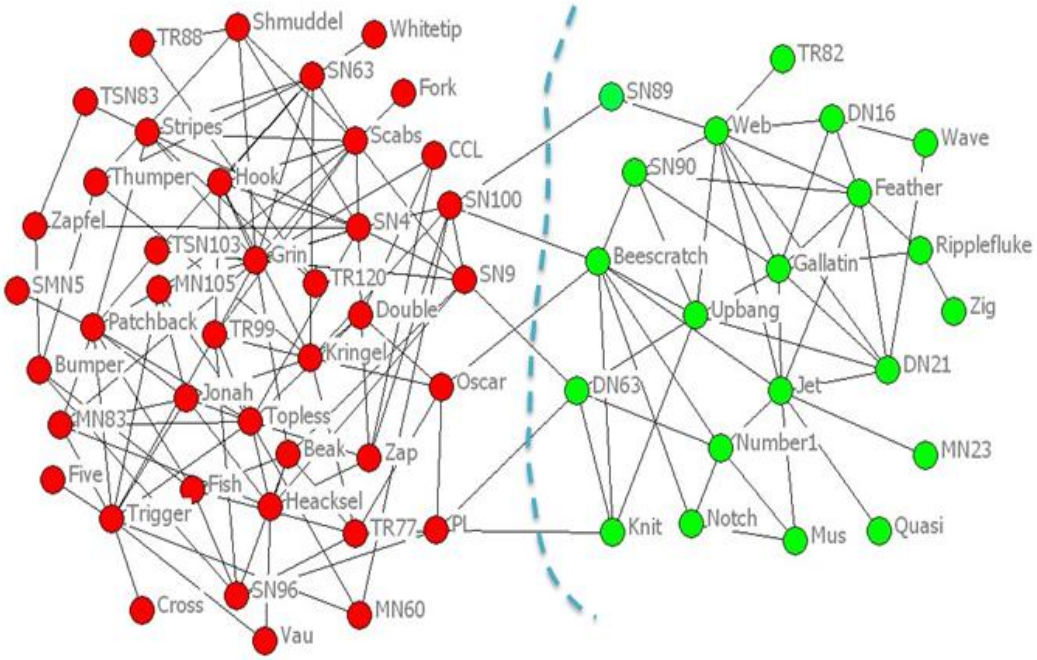

Figure 5.5: GN requires 6 iterations to discover the two communities in Dolphin social network.

5.3.5 Observations

One of the best algorithms for community discovery has been that proposed by Newman for detecting exact communities without disturbing internal links. But the algorithm is not scalable for huge graphs. Our enhanced algorithm is reducing by 40% of time complexity on average when compared to Newman algorithm for detecting exact communities. We summarize our implementational observations as follows. Note that we refer to an 'edge having a common neighbour' to mean that the endpoints of the edge have a common neighbour.

1. Remove edge with highest betweenness score same as GN algorithm.
2. Addtionally remove the other top edges with next highest betweenness score which have no common neighbours in the same iteration. For example as seen in the table: 5.6 for Santa Fe collaboration network, 78-64 is first removed and the second highest betweenness edge with no common neighbours 34-27 is also removed in the same iteration 1.
3. The top most highest betweenness edge is removed even if it is part of a triangle. After that see whether the next highest betweenness edge has common neighbour or not, See in table: 5.6 we remove the highest betweenness edge 64-41 and after that the next highest 65-41 and 63-41 edges are also removed, in the same iteration as they do not have common neighbours.
4. Note that, we do not remove even a second highest betweenness edge if it has a common neighbour. See in table 5.7, the edge $7-1$ is not removed in the same iteration even though it has highest betweenness score.
5. Proposed Enhanced-GN algorithm works well for sparse graphs, may be slower for dense graphs. Overall the complexity is reduced when compared to Newman algorithm.

We consider the output of GN algorithm as the ground truth against which we compare our results, if i is the number of iterations taken by our algorithm and N is that taken by GN then % of saving $= \frac{N-i}{N} * 100$. Let j be the number of nodes that are clustered exactly the same as GN algorithm then we measure accuracy $= \frac{j}{n} * 100$, n is total number of nodes in G.

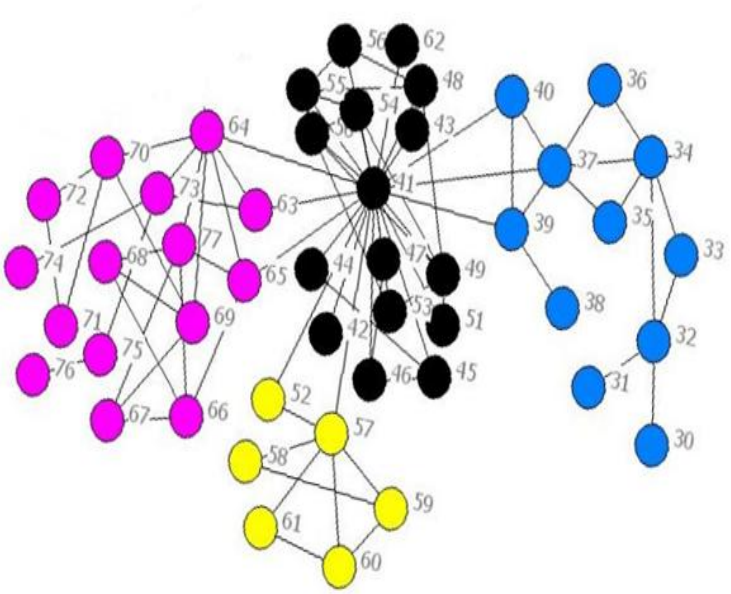

Figure 5.6: A subgraph of Santa Fe collaboration graph is shown. The edges 64-41, 63-41 and 65-41 are removed in the same iteration by Enhanced-GN algorithm.

Dataset	# iteration for GN	# iteration Enhanced-GN	% Saving	% Accuracy
Zachary karate club	11	5	54.00	97.05
Santa Fe-collaboration	17	13	23.00	100
Dolphin social network	6	3	50.00	100
Netscience network	11	4	53.80	95
GP author collaborat.	13	4	65.60	98
Facebook network	9	6	33.33	100
Sparce network	14	6	57.10	96.3

Table 5.8: Comparison of efficiency with GN at a threshold $t = 0.3$ for different datasets.

5.3.6 Threshold analysis

The algorithm is run for different thresholds of derivation varying between 0.2 to 0.5. The results obtained regarding % of iterations saved and accuracy of community composition are tabulated in table: 5.9. It can be seen that the value 0.3 is the least of the thresholds that can be chosen for optimal performance.

	Results obtained by varying the threshold									
	0.2		0.25		0.3		0.35		0.4	
Datasets	% Saving	% Accuracy	% Saving	% Accu.	% Saving	% Accu.	% Saving	% Accu.	% Saving	% Accu.
Karate club	45	97.05	45	97.05	54	97.05	54	97.05	54	97.05
Santa Fe.	17	98.30	7	98.30	23	100	23	100	23	100
Dolphin net.	50	100	50	100	50	100	50	100	50	100

Table 5.9: % of saving and % of accuracy at different deviation thresholds are given for the bench mark data sets.

5.3.7 Implementation of Enhanced-GN on real world datasets

In Chapter 3, the tertiary attachment model was discussed for several real world data sets like Genetic Programming(GP), Netscience and Facebook(FB). We run our Enhanced-GN algorithm to discover communities for these data sets and additionally we consider a sparse network from FB with 135 nodes and 139 edges with no triangles figure: 4.30. In the table: 5.10 we tabulate the results of Enhanced-GN algorithm on these datasets. It can be seen that the algorithm achieves a saving of 56% for Netscience, 63% for GP, 33% for FB and 57% for the sparse network. The accuracy is preserved as can be seen in the table: 5.11.

Dataset	Nodes	Average deg.	Avg. clus coeff.	Avg. path len.
GP collaboration	98	2.87	0.269	3.42
Netscience network	109	2.85	0.154	6.69
Facebook network	121	2.91	0.343	5.05
Sparse network	135	2.06	0.000	5.76

Table 5.10: Statistics of the real world benchmark data sets.

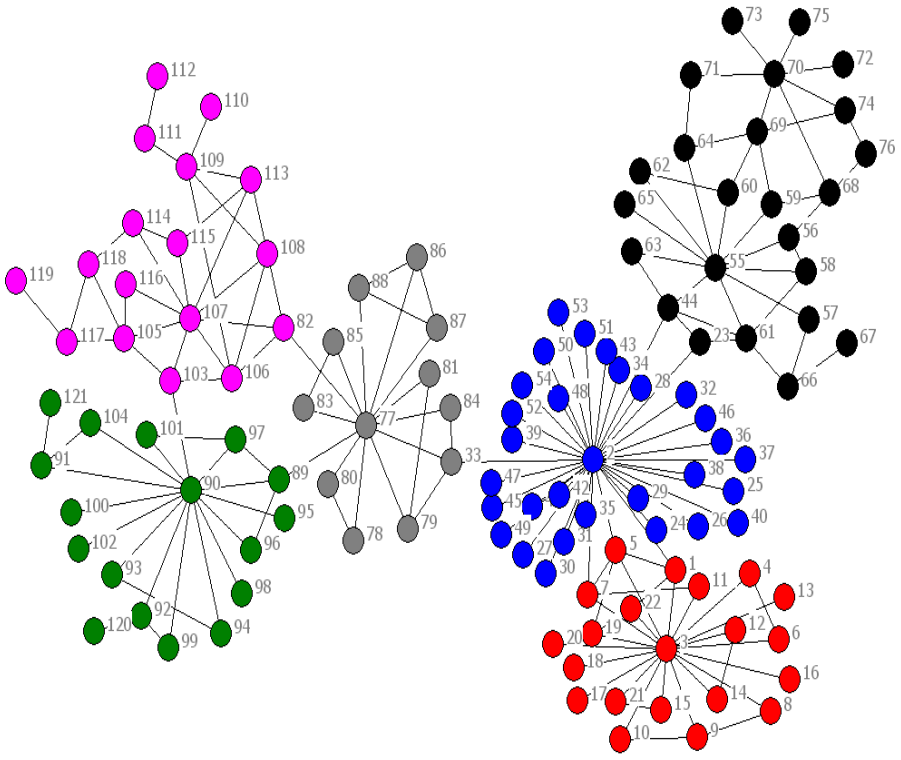

Figure 5.7: A snapshot of Facebook users of size 121 from year 2007.

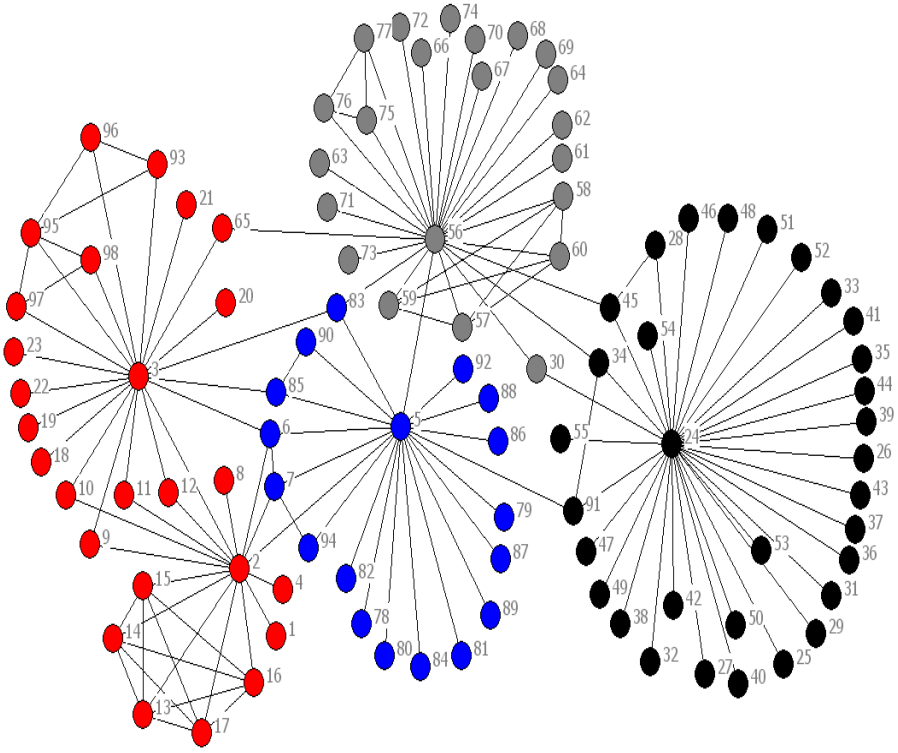

Figure 5.8: A snapshot of GP authors dataset of size 98 from a component.

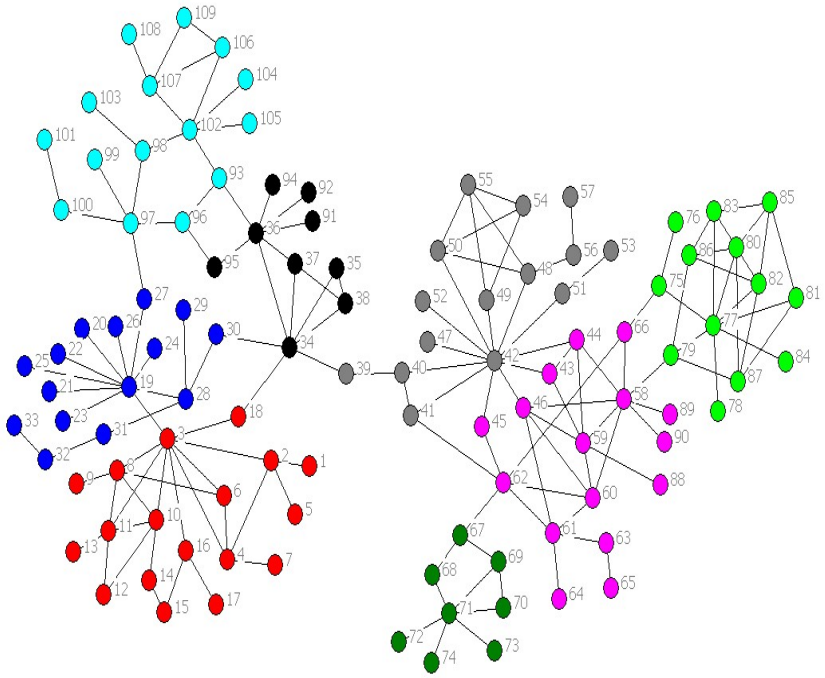

Figure 5.9: A snapshot of Netscience authors dataset of size 109 from a component.

Datasets	% Saving	% Accuracy
GP academic network	65.6	98
Netscience network	53.8	95
Facebook-1	33.33	100
Facebook-2 (Sparse)	57.00	96.3

Table 5.11: Table with different threshold of different datasets, our proposed algorithm sonsidered with 0.3% for all datasets.

5.3.8 Observations

1. Enhanced-GN algorithm performs quite significantly with respect to speedening up the GN algorithm on real world data sets also.
2. When the networks have less number of triangles, like for Netscience and Facebook(Sparse)(seen by their low values of clustering coefficient), we observe that the algorithm removes some of the internal edges which affects the correctness of the algorithm.
3. We also see that in these graphs, the removal of edges with first and second highest scores may create isolated nodes when these nodes are not joined by a triangle, thus again affecting accuracy.
4. We feel that if the graph has a minimum clustering coefficient of 0.25, our algorithm performs well in saving the number of iterations as well grouping the nodes correctly among communities.

5.4 Applications of Enhanced-GN

We propose to solve the influence maximization problem of social networks by retrieving influential nodes within communities using the Enhanced-GN algorithm. The edge-betweenness score used by GN algorithm has never been utilized for this purpose in the literature. We believe that this score can be used to indicate the amount of influence that a node has on its neighbours.

5.4.1 Influence maximization

Influence maximization is a problem of selection of a subset of nodes which maximizes influence in a social network. Recently Charu C. Aggarwal and Yan (2011) proposed a stochastic information flow model and an algorithm called RankedReplace

to retrieve influential nodes. Though RankedReplace algorithm gives better information spread, it becomes very slow for large data sets.

RankedReplace algorithm is slow since it has a heavy initialization step involving steadyStateSpread procedure. There have been attempts to speed up RankedReplace algorithm using different heuristics at the initialization step (Mohammed Mustafa, M.Tech thesis, University of Hyderabad). We propose that community discovery may help in choosing influence nodes for initialization and then carry out RankedReplace algorithm as it is for maximizing information spread. For details regarding steadyStateSpread procedure, refer to Charu C. Aggarwal and Yan (2011). The idea is as follows:

Step 1: Discover communities using Enhanced-GN algorithm

Step 2: Take the max-degree nodes from each community to constitute initial set in the RankedReplace algorithm.

Step 3: Execute the inner loop of RankedReplace algorithm

We define flow-value for a node i as the maximum edge betweenness score among all the edges at i. That is,

$$fv(i) = \max_{j \in N(i)} C'_{\mathrm{B}}(i,j)$$

Algorithm 3 Community-RR(P:probMatrix, k:number of communities, r:factor of nodes from a community, m:maxiterations)

procedure (Influence maximization)
  $S = \phi$
  Find communities C of P using Enhanced-GN algorithm
  for each community **do**
    $S = S \cup$ {top r nodes of maximum degree in C}
  end for
  S: Initial set of kr authority nodes
  Arrange nodes in $(V - S)$ in descending order of *fv(i)*.
  for each node i in $(V - S)$ in descending order **do**
    sort the list S in ascending order of *fv(i)*
    pick the first element of sorted list S which is such that replacing i with it increases value of *steadyStateSpread(S, P)*
    if no replacement has occurred in the last m consecutive iterations, ***return(S)*** and terminate.
  end for
  Return (S)
end procedure

5.4.2 Results

We run the algorithm $Community - RR$ on two data sets Zachary and a portion of Netscience. The results tabulated in table:5.12 clearly show that this algorithm speeds up RankedReplace algorithm by reducing the number of replacements in each step. The results are also compared with DegreeDiscount heuristic (RRDD) as proposed by Wang and Yang (2009).

	Number of replacements				Influence spread		
Datasets	Size	RR	RRDD	Commu-RR	RR	RRDD	Commu-RR
Netscience	4	10	17	10	$6.20 * e^{01}$	$6.20 * e^{01}$	$6.09 * e^{01}$
	5	14	26	15	$6.25 * e^{01}$	$6.25 * e^{01}$	$6.21 * e^{01}$
	6	18	19	19	$6.41 * e^{01}$	$6.32 * e^{01}$	$6.39 * e^{01}$
	7	17	23	19	$6.48 * e^{01}$	$6.47 * e^{01}$	$6.39 * e^{01}$
Zachary	2	0	19	0	$2.49 * e^{01}$	$2.49 * e^{01}$	$2.36 * e^{01}$

Table 5.12: Results show that number of replacements taken by Commu-RR is very close to that of RR, without significant reduction in influence spread.

The algorithm needs to be implemented on large bench mark data sets for further validation. It is clear intuitively that if influential nodes within communities are given as initial nodes to the influence maximization algorithm, the influence will spread very quickly within the community. As each of the communities is covered by at least one node, the algorithm is faster with as few replacements as the algorithm of Charu C. Aggarwal and Yan (2011). It is interesting to note that the proposed algorithm performs better than the heuristic of DegreeDiscount in terms of speed. Certainly, the amount of influence spread needs to be improved by the proposed algorithm which is part of the future work.

5.5 Conclusions

Community discovery is an NP-Hard problem. One of the most popular and efficient heuristic algorithms for the problem is that of (GN)Girvan and Newman (2002). This algorithm becomes very slow for large sized networks. In this context, we propose an additional heuristic to enhance the speed of GN algorithm. The algorithm is tested on the standard bench mark data sets and is shown to give a significant improvement

on speed without suffering in accuracy. The proposed Enhanced-GN algorithm is also tested on small component subsets of real-world networks. The algorithm performs quite well on these data sets also.

One of the limitations of the algorithm seems to be for very sparse networks, in particular, networks whose clustering coefficient is zero. We see that when a network does not possess any triangles, the correctness of the algorithm suffers. In terms of speed, the worst case for the algorithm is when the network is very dense.

Another problem that we addressed in this chapter is that of influence maximization in social networks. We use the idea that if communities are extracted in a social network and consider high degree nodes within communities, they may prove to be efficient seeds for the RankedReplace algorithm of Charu C. Aggarwal and Yan (2011). This idea along with the influence being calculated using the edge-betweenness score works very well. We test the proposed influence maximization algorithm on a few bench mark data sets. It speeds up the existing algorithms quite well. The algorithm needs to be tested on larger bench mark data sets for further validation.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

Many of the network evolution models in the literature use the property of transitivity or formation of triads for network growth. One of the simplest models, from computational complexity point of view is that of R.Toivonen *et al.* (2006). Apart from directly linking to a given node, they propose a mechanism, which we name as secondary attachment, to depict friend-of-friend connections to grow the network. We observe that in real world scenarios, not only secondary but tertiary attachments(friend-of-friend-of-friend) are also quite common. Hence in the first part of the thesis we investigate this idea in a thorough manner. Tertiary attachment model brings in ties of the type friend-of-friend (f-of-f) as well as f-of-f-of-f, tertiary connections.

6.1 Conclusions

We propose a tertiary attachment(TA) model with stochastic distributions m_r, m_s and m_t from which to choose primary, secondary and tertiary contacts respectively at each time step. We prove mathematically by deriving the rate equation for rate of change of average degree, that the model does evolve a scale free network since the probability density function of the degree distribution of nodes of degree k is shown as $P(k) \sim k^{-\gamma}$ with $\gamma = \left(3 + \frac{2}{m_s+2m_s m_t}\right)$ leading to $3 < \gamma < \infty$. The rate equation for clustering coefficient shows that $c(k)$ evolves as $\frac{\log k}{k}$ with a slightly larger rate than the model of R.Toivonen *et al.* (2006). We show that the proposed TA model is important for social networks since it exhibits crucial properties of social networks like short average path length to signify small world property, high average clustering coefficient and associativity to show strong community structure and skewed degree distribution to show the scale-free property of the network.

We assess the applicability of the tertiary attachment model in the context of two different kinds of of social networks, namely, collaboration networks and friendship

networks. One knows that these two networks grow very differently since it takes relatively much longer to establish a research collaboration than to link to a friend in a typical friendship network. This model deepens the perspective for studying structure formation and evolutionary mechanisms. Genetic Programming(GP) data set and Facebook(FB) data set for which time stamps are available are chosen for this study. We find that the secondary and tertiary attachments together amount to nearly 40% of the total edges in the GP network and 45% in FB network, thus validating the basis of our model. We carried out extensive analysis by defining different types of triangles $T1, T2, T3$ and $T4$, with $T2$ and $T3$ roughly corresponding to secondary and tertiary connection formation. We find that the density of $T3$ and $T4$ amount to 11% of the total triangles in GP and 37% in FB which is quite high validating the basis for our model. Algorithms need to be worked out for link prediction incorporating these ideas, which can then prove the predictive capability of the model.

Further, we show that the rate equations derived for the TA model, when simulated with different parameter distributions, match very closely with the rate of change of degree and clustering coefficient in GP. The data sets of FB do not give good results which needs to be investigated further.

In the last part of the thesis, we tackle one of the important problems for social networks called the Community discovery. Girvan and Newman (GN)Girvan and Newman (2002) provide a classical approach to this problem using the measure of edge-betweenness score. This algorithm may not be scalable for large data sets. In this context, we propose a heuristic that enhances the speed of GN algorithm. We test our algorithm on data sets like Zachary karate club, Santa Fe collaboration network and Dolphin network which are some of the standard benchmark data sets for community discovery. This heuristic algorithm called Enhanced-GN algorithm shows significant improvement over the GN algorithm, along with retaining the accuracy. Further, the algorithm is tested on additional real-world data sets by choosing portions from GP, FB and Netscience We show that Enhanced-GN algorithm performs quite well by speeding up the GN algorithm on real world data sets also. Extensive experimentation on larger data sets is necessary for further val-

idation. An application of Enhanced-GN algorithm is shown for solving influence maximization problem. This algorithm opens up new ideas for solving this problem and needs to be tested on large data sets to test its efficacy.

6.2 Future directions

It has been shown TA model simulates collaboration networks quite closely and its applicability to Facebook data set is not so clear. There is a great need to generate larger benchmark data sets with time stamp, in order to assess the proposed model as well as evolve new generating mechanisms. It will be interesting to study the current Facebook network which has immense growth to see the adaptability of our model. We have to make a note here that our initial experiments with higher order attachments showed that almost completely connected graphs are produced which are not useful. Hence tertiary models seem to be optimal.

Analysis regarding the different kind of triangles which roughly correspond to primary, secondary and tertiary connections should be further studied from the point of view of predictive capability of the model. Using these ideas, we need to investigate further that if a part of the network is known, then with appropriate probabilistic analysis, can the model predict secondary or tertiary attachments between two given nodes.

Most of the existing models include edge deletion as part of their mechanismKumpula *et al.* (2007) to inhibit the network growth. It will be interesting to include a similar rule in our TA model algorithm to see the kind of networks that evolve. Random link removal is supposed to reinforce community structure and hence this mechanism may ensure stronger community structure for our model.

It is important to carry out stability analysis for the emergent network. That is, to assess if small changes in the optimal parameter set will lead to large changes in the network. A limited analysis has been carried out in the thesis, by varying the parameter distributions and it is observed that the degree distribution and average clustering coefficient do not change significantly. In fact, we find that the general

trend of the average statistics is preserved among all the parameter sets considered. A deeper analysis needs to be carried out both for the secondary attachment model of R.Toivonen *et al.* (2006) as well as TA model for assessment of stability.

We believe that there is much scope for applying betweenness centrality measure to various applications. The recent work of Ufimtsev and Bhowmick (2013) that proposes an efficient method for retrieving top few highest centrality nodes will give a boost to many applications in this area. This work will directly and positively impact the efficiency of our Enhanced-GN algorithm and hence also our algorithm for influence maximization. Of course, extensive validation of our proposed algorithm for larger benchmark datasets needs to be carried out which is part of ongoing work.

REFERENCES

1. **Airoldi, E. M.**, *Bayesian Mixed-membership Models of Complex and Evolving Networks*. Carnegie Mellon University, 2006.

2. **Alba, R. D.** (1973). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, **3**, 113–126.

3. **Arabie, P.**, **S. A. Boorman**, and **P. R. Levit** (1978). Constructing blockmodels: How and why. *Journal of Mathematical Psychology*, **17(1)**, 21–63.

4. **Ball, B.** and **M. E. J. Newman** (2012). Friendship networks and social status. *CoRR*, **abs/1205.6822**.

5. **Banks, D. L.** and **K. M. Carley** (1996). Models for network evolution. *Journal of Mathematical Sociology*, 173–196.

6. **Barabasi, A. L.** and **R. Albert** (1999). Emergence of scaling in random networks. *Science (New York, N.Y.)*, **286**(5439), 509–512.

7. **Blitzstein, J.** and **P. Diaconis** (2006). A sequential importance sampling algorithm for generating random graphs with prescribed degrees.

8. **BollobÃąs, B.**, **S. Janson**, and **O. Riordan** (2007). The phase transition in inhomogeneous random graphs. *Random Struct. Algorithms*, **31**, 3–122.

9. **Brandes, U.** (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, **25**, 163–177.

10. **Buscarino, A.**, **M. Frasca**, **L. Fortuna**, and **A. Fiore** (2012). A new model for growing social networks. *Systems Journal, IEEE*, **6**(3), 531–538.

11. **Capocci, A.**, **V. D. P. Servedio**, **F. Colaiori**, **L. S. Buriol**, **D. Donato**, **S. Leonardi**, and **G. Caldarelli** (2006). Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, **74**(3).

12. **Catanese, S.**, **P. D. Meo**, **milio Ferrara**, and **G. Fiumara** (2011). Analyzing the facebook friendship graph. *CoRR*, **abs/1011.5168**.

13. **Catanzaro, M.**, **G. Caldarelli**, and **L. Pietronero** (2004). Assortative model for social networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, **70**(3).

14. **Charu C. Aggarwal, A. K.** and **X. Yan** (2011). On flow authority discovery in social networks. *Proc. of the Eleventh SIAM, SDM*.

15. **Chen, D.**, **M. Shang**, **Z. Lv**, and **Y. Fu** (2010). Mechanics and its applications : Detecting overlapping communities of weighted networks via a local algorithm. *Physica A*, **389**, 4177–4187.

16. **Chung, F.** and **L. Lu**, *Complex Graphs and Networks (Cbms Regional Conference series in Mathematics)*. American Mathematical Society, Boston, MA, USA, 2006.

17. **Clauset, A.** (2005). Finding local community structure in networks. *Phys. Rev. E*, **72**, 026–132.

18. **Dongen, S.**, A cluster algorithm for graphs. CWI (Centre for Mathematics and Computer Science), Amsterdam, Netherlands, 2000.

19. **Doreian, P.**, **V. Batagelj**, and **A. Ferligoj** (2004). Generalized blockmodeling of two-mode network data. *Social Networks*, **26**(1), 29–53.

20. **Du, N.**, **B. Wang**, and **B. Wu** (2008). Community detection in complex networks. *J. Comput. Sci. Technol.*, **23**(4), 672–683.

21. **Erdos, P.** and **A. Renyi** (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.*, **5**, 17–61.

22. **Fenner, T.**, **M. Levene**, **G. Loizou**, and **G. Roussos** (2007). A stochastic evolutionary growth model for social networks. *Comput. Netw.*, **51**(16), 4586–4595.

23. **Ferrara, E.** (2011). A large-scale community structure analysis in facebook. *CoRR*, **abs/1106.2503**.

24. **Ferrara, E.** and **G. Fiumara** (2012). Topological features of online social networks. *CoRR*, **abs/1202.0331**.

25. **Ferrara, E.**, **P. D. Meo**, **G. Fiumara**, and **A. Provetti** (2012). The role of strong and weak ties in facebook: a community structure perspective. *CoRR*, **abs/1203.0535**.

26. **Fienberg, S.**, *The Analysis of Cross-Classified Categorical Data*. Springer, 2007, 2nd edition.

27. **Flaxman, A. D.**, **A. M. Frieze**, and **J. Vera**, A geometric preferential attachment model of networks. *In Algorithms and Models for the Web-Graph*. Springer, 2004.

28. **Frank, O.** and **D. Strauss** (1986). Markov graphs. *Journal of the American Statistical Association*, **81**(395), 832–842.

29. **Freeman, L.** (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–41.

30. **Fu, X.**, **C. WANG**, **Z. WANG**, and **Z. MING** (2012). Scalable community discovery based on threshold random walk. *Journal Of Computational Information Systems*, **8**(21), 8953–8960.

31. **Gilbert, E.** (1959). Random graphs. *Annals of Mathematical Statistics*, **30**, 1141–1144.

32. **Girvan, M.** and **M. E. J. Newman** (2002). Community structure in social and biological networks. *PNAS*, **99**(12), 7821–7826.

33. **Gjoka, M.**, **M. Kurant**, **C. T. Butts**, and **A. Markopoulou**, Walking in facebook: A case study of unbiased sampling of osns. *In Proceedings of the 29th Conference on Information Communications*, INFOCOM. IEEE Press, Piscataway, NJ, USA, 2010.

34. **Gjoka, M.**, **M. Sirivianos**, **A. Markopoulou**, and **X. Yang**, *Poking Facebook: Characterization of OSN Applications*. Seattle, WA, 2008.

35. **Goodreau, S. M.** (2007). Advances in exponential random graph p* models applied to a large social network. *SocNet*. Références; intro de revue spéciale.

36. **Granovetter, M.** (1973). The strength of weak ties. *American journal of sociology*, 1360–1380.

37. **Gregory, S.**, An algorithm to find overlapping community structure in networks. *In ECPP*, PKDD. Springer-Verlag, 2007.

38. **Gregory, S.**, A fast algorithm to find overlapping communities in networks. *In ECML-Part I*, PKDD. Springer-Verlag, Berlin, Heidelberg, 2008.

39. **Handcock, M.** (2003). Assessing degeneracy in statistical models of social networks. Technical Report Working Paper 39, University of Washington.

40. **Handcock, M. S.**, **D. R. Hunter**, **C. T. Butts**, **S. M. Goodreau**, and **M. Morris** (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, **24**.

41. **Hanneke, S.** and **E. P. Xing**, Discrete temporal models of social networks. *In CSNA*, ICML. Springer-Verlag, Berlin, Heidelberg, 2007.

42. **Holland, P.** and **S. Leinhardt** (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, **5**, 5–20.

43. **Holland, P. W.** and **S. Leinhardt** (1981). An exponential family of probability distributions for directed graphs. with comments by Ronald L. Breiger. *Journal of the American Statistical Association*, **76**(373), 33–65.

44. **Holme, P.**, **C. Edling**, and **F. Liljeros** (2004). Structure and time evolution of an internet dating community. *Social Networks*, **26**(2), 155–174.

45. **Hu, H.** and **X. Wang** (2012). How people make friends in social networking sites a microscopic perspective. *Physica A: SMA*, **391**(4), 1877–1886.

46. **Ishida, K.**, **F. Toriumi**, and **K. Ishii**, Proposal for a growth model of social network service. *In Web Intelligence*. IEEE, 2008.

47. **Jeong, H.**, **Z. Neda**, and **A. BarabÃąsi** (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters*, **61**(4), 567–572.

48. **J.M., R.** (2002). Simple methods for simulating sociomatrices with given marginal totals. *Social networks*, 273–283.

49. **Kanovsky, I.** (2010). Small world models for social network algorithms testing. *Procedia Computer Science*, **1**(1), 2341–2344.

50. **Kernighan, B. W.** and **S. Lin** (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, **49**(2), 291–307.

51. **Krapivsky, P. L.** and **S. Redner** (2001). Organization of growing random networks. *Physical Review E*, **63**, 066–123.

52. **Kumpula, J. M.**, **J. P. Onnela**, **J. Saram a ki**, **K. Kaski**, and **J. a. n. Kert e sz** (2007). Emergence of Communities in Weighted Networks. *Physical Review Letters*, **99**, 228701+.

53. **Lee, M.-J.**, **J. Lee**, **J. Y. Park**, **R. H. Choi**, and **C.-W. Chung** (2012). Qube: a quick algorithm for updating betweenness centrality, 351–360.

54. **Lewis, K.**, **J. Kaufman**, **M. Gonzalez**, **A. Wimmer**, and **N. Christakis** (2008). Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, **30**(4), 330–342.

55. **Li, K.**, **X. Gong**, **S. Guan**, and **C. Lai** (2012). Efficient algorithm based on neighborhood overlap for community identification in complex networks. *Physica A: Statistical Mechanics and its Applications*, **391(4)**, 1788–1796.

56. **Liu, H.-T.**, **D. Pei**, and **Y. Wu** (2012). A novel evolution model of collaboration network based on scale-free network. *Springer Berlin Heidelberg*, 148–155.

57. **Luce, R.**, **J. Macy**, and **R. Tagiuri** (1955). A statistical model for relational analysis. *Psychometrika*, **20**(4), 319–327.

58. **Luce, R. D.** (1950). Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, **15**, 169–190.

59. **Luce, R. D.** and **A. D. Perry** (1949). A method of matrix analysis of group structure. *Prychometrika*.

60. **Luthi, L.**, **M. Tomassini**, **M. Giacobini**, and **W. B. Langdon**, The genetic programming collaboration network and its communities. *In Proceedings of the 9th annual conference on Genetic and evolutionary computation*, volume 2. ACM Press, 2007.

61. **M, A. E.** (2007). Getting started in probabilistic graphical models. *PLoS Comput Biol*, **3**, 252.

62. **Maria, A. E.** (2005). Model based clustering for social networks. *Journal of the Royal Statistical Society*, (46).

63. **Milgram** and **Stanley** (1967). The small world problem. *Psychology Today*, **1**(1), 61–67.

64. **Mishra, N.**, **R. Schreiber**, **I. Stanton**, and **R. E. Tarjan** (2008). Finding strongly knit clusters in social networks. *Internet Mathematics*, **5**, 155–174.

65. **Moreno, J.**, *Who shall survive? Nervous and mental disease publishing company*. Washington,D.C, 1934.

66. **Newman, . M. E. J.** (2003). The structure and function of complex networks. *SIAM*, **45**(12), 167–256.

67. **Newman, M.** (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, **64**(2), 025–102.

68. **Newman, M. E.** (2006). Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, **103**, 8577–8582.

69. **Newman, M. E. J.** (1998). The structure of scientific collaboration networks. *Proc. - National acady of sciences, USA*, **98**(ISSN 0027-8424), 404–409.

70. **Newman, M. E. J.** (2000). Models of the small world. *J. Stat. Phys*, 819–841.

71. **Newman, M. E. J.** (2004*a*). Coauthorship networks and patterns of scientific collaboration. *Proc. - National acady of sciences, USA*, **101**(ISSN 0027-8424), 5200–5205.

72. **Newman, M. E. J.** (2004*b*). Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, **38**, 321–330. ISSN 1434-6028.

73. **Newman, M. E. J.**, **D. J. Watts**, and **S. H. Strogatz** (2002). Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, **99**, 2566–2572.

74. **Radicchi, F.**, **C. Castellano**, **F. Cecconi**, **V. Loreto**, and **D. Parisi** (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, **101**(9), 2658.

75. **Radner, R.** and **A. Tritter** (1954). Communication in networks.

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

77. **R.Toivonen**, **J.-P. Onnela**, **J. Saramaki**, **J. Hyvonen**, and **K. Kaski** (2006). A model for social networks. *Physica A: 371*, 851–860.

78. **Shalizi, C. R.**, **M. F. Camperi**, **Klinkner**, and **K. Lisa** (2007). Discovering functional communities in dynamical networks. *Springer*, **4503**, 140–157.

79. **Snijders, T.**, Models for longitudinal network data. *In* **P. Carrington**, **J. Scott**, and **S. Wasserman** (eds.), *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005, 215–247.

80. **Snijders, T. A. B.** (2006). Statistical methods for network dynamics. *Journal of Mathematical Sociology*, **Statistical Society**, 281–296.

81. **Spilerman, S.** (1966). Structural analysis and the generation of sociograms. *Behav Sci*, **11**(4), 312–8.

82. **Szabo, G.**, **M. Alava**, and **Kertsz** (2003). Structural transitions in scale-free networks. *Physical Review E*, **67**, 056–102.

83. **Szczepanski, P. L.**, **T. Michalak**, and **T. Rahwan** (2012). A new approach to betweenness centrality based on the shapley value. *IFAAMS*, **1**, 239–246.

84. **Tang, J.**, **J. Zhang**, **L. Yao**, **J. Li**, **L. Zhang**, and **Z. Su**, Arnetminer: Extraction and mining of academic social networks. KDD. ACM, New York, USA, 2008.

85. **Toivonen, R.**, **L. Kovanen**, **K. Mikko**, **J.-P. Onnela**, **J. Saramaki**, and **K. Kaski** (2009). A comparative study of social network models: network evolution models and nodal attribute models. *Elsevier*, **31**, 240–254.

86. **Tomassini, M.** and **L. Luthi** (2007). Empirical analysis of the evolution of a scientific collaboration network. *Physica A*, 750–764.

87. **Traud, A. L.**, **P. J. Mucha**, and **M. A. Porter** (2011). Social structure of facebook networks. *CoRR*, **abs/1102.2166**.

88. **Ufimtsev, V.** and **S. Bhowmick** (2013). Application of group testing in identifying high betweenness centrality vertices in complex networks. (ACM 978-1-4503-2322-2).

89. **van Duijn, M. A. J.**, **T. A. B. Snijders**, and **B. J. H. Zijlstra** (2004). P2: a random effects model with covariates for directed graphs. *Statistica neerlandica*, **58**, 234–254.

90. **Vázquez, A.** (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E*, **67**, 056104.

91. **Viswanath, B.**, **M. Alan**, **C. Meeyoung**, and **G. K. P.** (2009). On the evolution of user interaction in facebook. *ACM*, 37–42.

92. **Von Arb, M.**, **M. Bader**, **M. Kuhn**, and **R. Wattenhofer**, Veneta: Serverless friend-of-friend detection in mobile social networking. *In Wimob*. IEEE, 2008.

93. **Wang, M.**, **G. Yu**, and **D. Yu** (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A*, **387**(18), 4692–4698.

94. **Wang, W. C. Y.** and **S. Yang** (2009). Efficient influence maximization in social networks. *ACM SIGKDD*, 199–208.

95. **Wasserman, S.** and **K. Faust**, *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

96. **Wasserman, S.** and **P. Pattison** (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, **61**, 401–425. ISSN 0033-3123.

97. **Wasserman, S. S.** (1977). Stochastic models for directed graphs.

98. **Watts, D.** and **S. Strogatz** (1998). Collective dynamics of small world networks. *Nature*, **393**(6684), 440–442.

99. **Wei Deng, J.**, **K. ying Deng**, **Y. sheng Li**, and **Y. xing Li**, Study on evolution model and simulation based on social networks. *In ICNC*. IEEE, 2012.

100. **Yan, B.** and **S. Gregory** (2012). Detecting community structure in networks using edge prediction methods. *Journal of Statistical Mechanics: Theory and Experiment*, **2012**.

101. **Yule, G. U.** (1925). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philosophical transactions of the Royal Society of London*, **213**, 21–87.

102. **Zijlstra, H.**, **M. A. J. van Duijn**, and **T. A. B. Snijders** (2006). The multilevel p2 model: A random effects model for the analysis of multiple social networks. *Methodology*, **2(1)**, 42–47.

LIST OF PAPERS BASED ON THESIS

1. Sreedhar Bhukya and S.Durga Bhavani, Community Discovery in Social Networks, to be communicated.

2. Sreedhar Bhukya. A novel model for social networks, *Baltic Congress on Future Internet Communication (BCFIC),IEEE, 2011, pp 21-24.*

3. Sreedhar Bhukya. A social network model for academic collaboration, *Networked Digital Technologies (NDT), CCIS,1, Volume 136, Part 3, pp.203-211. (in DBLP)*

4. Sreedhar Bhukya. Discover academic experts in novel social network model, *Advances in Social Networks Analysis and Mining(ASON AM),IEEE, 2011, pp 696-700. (in DBLP)*

5. Sreedhar Bhukya. Information propagation on novel social network model, *Computer Networks and Intelligent Computing, Volume 157, Springer Berlin Heidelberg, 2011, pp 141-148.*

6. Sreedhar Bhukya. A novel social network model for forming relationships, *DEIS-2011, Communications in Computer and Information Science (CCIS), Volume 194, Springer Berlin Heidelberg, 2011, pp 287-295. (in DBLP)*