Models for the Generation of Heterogeneous Complex Networks

Bassant El-Sayed Youssef

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Scott F. Midkiff, Chair
Ing-Ray Chen
Luiz A. DaSilva
Hoda M. Hassan
Yiwei Thomas Hou
Mohamed Rizk Mohamed Rizk

June 8, 2015
Blacksburg, Virginia

Keywords: Complex networks, Mathematical modeling, Preferential attachment

Models for the Generation of Heterogeneous Complex Networks

Bassant El-Sayed Youssef

ABSTRACT

Complex networks are composed of a large number of interacting nodes. Examples of complex networks include the topology of the Internet, connections between websites or web pages in the World Wide Web (WWW), and connections between participants in social networks. Due to their ubiquity, modeling complex networks is important for answering many research questions that cannot be answered without a mathematical model. For example, mathematical models of complex networks can be used to find the most vulnerable nodes to protect during a virus attack in the Internet, to predict connections between websites in the WWW, or to find members of different communities in social networks. Researchers have analyzed complex networks and concluded that they are distinguished from other networks by four specific statistical properties. These four statistical properties are commonly known in this field as: (i) the small world effect, (ii) high average clustering coefficient, (iii) scale-free power law degree distribution, and (iv) emergence of community structure. These four statistical properties are further described later in this dissertation.

Most models used to generate complex networks attempt to produce networks with these statistical properties. Additionally, most of these network models generate homogeneous complex networks where all the network nodes are considered to have the same properties. Homogenous complex networks neglect the heterogeneous nature of the nodes in many complex networks. Moreover, some models proposed for generating heterogeneous complex networks are not general as they make specific assumptions about the properties of the network.Including heterogeneity in the connection algorithm of a model would make it more suitable for generating the subset of complex networks that exhibit selective linking. Additionally, all models proposed, to date, for generating heterogeneous complex networks do not preserve all four of the statistical properties of complex networks stated above. Thus, formulation of a model for the generation of general heterogeneous complex networks with characteristics that resemble as much as possible the statistical properties common to the real-world networks that have received attention from the research community is still an open research question.

In this work, we propose two new types of models to generate heterogeneous complex networks. First, we introduce the Integrated Attribute Similarity Model (IASM). IASM uses preferential attachment (PA) to connect nodes based on a similarity measure for node attributes combined with a node's structural popularity measure. IASM integrates the attribute similarity measure and a structural popularity measure in the computation of the connection function used to determine connections between each arriving (newly created) node and the existing (previously created or old) network nodes. IASM is also the first model known to assign an attribute vector having more than one element to each node, thus allowing different attributes per node in the generated complex network. Networks generated using IASM have a power law degree distribution and preserve the small world phenomenon. IASM models are enhanced to increase their clustering coefficient using a triad formation step (TFS). In a TFS, a node connects to the neighbor of the node to which it was previously connected through preferential attachment, thus forming a triad. The TFS increases the number of triads that are formed in the generated network which increases the network's average clustering coefficient.

We also introduce a second novel model, the Settling Node Adaptive Model (SNAM). SNAM reflects the heterogeneous nature of connection standard requirements for nodes. The connection standard requirements for a node refers to the values of attribute similarity and/or structural popularity of old node y that node new x would find acceptable in order to connect to node y. SNAM is novel in that such a node connection criterion is not included in any previous model for the generation of complex networks. SNAM is shown to be successful in preserving the power law degree distribution, the small world phenomenon, and the high clustering coefficient of complex networks.

Next, we implement a modification to the IASM and SNAM models that results in the emergence of community structure. Nodes are classified into classes according to their attribute values. The connection algorithm is modified to include the class similarity values between network nodes. This community structure model preserves the PL degree distribution, small world property, and does not affect average clustering coefficient values expected from both IASM and SNAM. Additionally, the model exhibits the presence of community structure having most of the connections made between nodes belonging to the same class with only a small percent of the connections made between nodes of different classes.

We perform a mathematical analysis of IASM and SNAM to study the degree distribution for networks generated by both models. This mathematical analysis shows that networks generated by both models have a power law degree distribution.

Finally, we completed a case study to illustrate the potential value of our research on the modeling of heterogeneous complex networks. This case study was performed on a Facebook dataset. The case study shows that SNAM, with some modifications to the connection algorithm, is capable of generating a network with almost the same characteristics as found for the original dataset. The case study provides insight on how the flexibility of SNAM's connection algorithm can be an advantage that makes SNAM capable of generating networks with different statistical properties.

Ideas for future research areas include studying the effect of using eigenvector centrality, instead of degree centrality, on the emergence of community structure in IASM; using the node index as an indication for its order of arrival to the network and distributing added connections fairly among network nodes along the life of the generated network; experimenting with the nature of attributes to generate a more comprehensive model; and using time sensitive attributes in the models, where the attribute can change its value with time.

Acknowledgements

First and above all, I praise God, the almighty for providing me the opportunity and granting me the capability to proceed successfully. This dissertation appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

I would like to express my special deep appreciation and hearty thanks to my advisor Professor Dr. Scott Midkiff .He has been a tremendous mentor always guiding me to structure and organize my ideas, setting high standards to achieve the best quality for my work, and enhancing my skills to grow as a good researcher, bearing my best interest in mind. I would like to thank him for his continuous encouragement. His advice on my research has been priceless. I really do appreciate his help, patience and the very valuable time he has provided me.

I would like to also thank my second advisor Professor Dr. Mohahmed Rizk. He has been ready with brilliant ideas, honest advice and encouraging words whenever I needed them; I enjoyed working with him, for the second time after my M.Sc. degree.

I would also like to thank all my committee members for their sincerity and useful discussions even at hardship. I also want to thank all my committee members for letting my defense be an enjoyable moment, and for their brilliant comments and suggestions.

Last but not least, I want to thank my loving and caring family. Words cannot express how grateful I am to my mother, and father for all of their sacrifices made on my behalf. Their prayer for me was what sustained me thus far. Their trust in me is what kept pushing me forward during this journey. I would like express appreciation to my sister Dr. Amira for the encouragement and support I received throughout the research work.

I would also like to thank all of my friends who supported me, and incented me to strive towards my goal. Special thanks to Mohamed Magdy and Ahmed Ibrahim for moral support to me throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

Chapter 1. Introduction

1.1. Introduction

Complex networks are ubiquitous in many areas. The Internet, the World Wide Web (WWW), social networks, food web (or food chain) networks, and many other networks are complex networks [1, 2]. Researchers have analysed these real-world complex networks, which has led to the discovery of their distinct statistical properties and behavioral patterns [3, 4]. The analysis of real-world complex networks has shown that most such networks possess four statistical properties. These four statistical properties are: (i) scale-free power-law degree distribution; (ii) small average path length, or small world phenomenon; (iii) high average clustering coefficient; and (iv) emergence of community structure [5, 6]. Devising a mathematical model for complex networks can aid in making decisions about the management of such networks and help in allocating their resources. There have been many attempts to find mathematical models that can faithfully generate networks that mimic real-world complex networks. Most models of complex networks that have been proposed do not result in all four of the common statistical properties of complex networks [7]. Since, nodes often have different characteristics from each other, they are heterogeneous. Most models for complex networks assume that nodes have the same characteristics, i.e. that nodes are homogeneous. Thus, these models neglect the heterogeneous nature of network nodes. Including heterogeneity in a generation model would make it more suitable for complex networks exhibiting assortative mixing. (Assortative mixing is defined in Section 3.1.). Moreover, even the models that have been proposed for heterogeneous complex networks do not integrate the heterogeneity of nodes with other structural properties of the network in the analysis and in the algorithms for generating such networks. Also, many existing models are specific for the generation of certain types of complex networks and, thus, are not general. Therefore, finding a faithful general heterogeneous complex network model for networks with assortative mixing that preserves the statistical properties of real-world complex network is still an open research question.

1.2. Contributions

In this research, we propose general mathematical models that are able to reflect the four statistical properties of complex networks, as described above. Moreover, our proposed models

consider node heterogeneity, a factor that, we claim, is unaddressed in most existing models of complex networks [4, 5].

We identify two types of node heterogeneity, heterogeneity of node characteristics and heterogeneity of node connection standards. Heterogeneity of node characteristics reflects the different properties or attributes of different network nodes. Heterogeneity of node connection standards reflects the difference in each node's requirements to make a connection to another given node.

The research contributions of this research is summarized as follows.

1) We account for the heterogeneity of node characteristics in a graph-theoretic model by incorporating node attributes as one of the elements defining a network graph. Accordingly, our model defines the network graph, G, as a set of three elements, $G = \{V, E, A\}$, where V is the set of nodes or vertices in the network, E is the set of links or edges, and A is the set of attribute-vectors assigned to each network node. The length of each vector in A is generally more than unity and not restricted to unity as in previously proposed heterogeneous generation models for complex networks.

2) Based on contribution (1) above, we propose the Integrated Attribute Similarity Model (IASM) for generating heterogeneous complex networks. IASM is based on the preferential attachment connection algorithm for generation of networks as proposed by Barbási and Albert [3]. However, IASM incorporates the heterogeneous nature of nodes by integrating attribute similarity with the structural popularity measure within the connection function, CF, used for the preferential attachment algorithm. Attribute similarity is used to assess the similarity or compatibility between the attributes of both nodes to be connected. In contrast, structural popularity measures the popularity of the old node based on its current connections. Structural popularity can be based on the number of the node's first degree connections (degree centrality) or it can consider higher degree connections (eigenvector centrality). Two models are proposed that use two different measures for node structural popularity. The first model, IASM_A, uses the nodes' normalized degree, while the second model, IASM_B, uses eigenvector centrality as a more accurate structural popularity measure. To increase the average clustering coefficient for networks generated by IASM, triad formation was added to IASM. The

triad formation step entails making an additional second degree connection after making an initial connection in the network based on preferential attachment, thus forming a triad. Increasing the number of triads (triangles) increases the value of the average clustering coefficient.

3) We introduce the idea of a heterogeneous node connection standard as a criterion for connecting new nodes during network growth. The connection standard of the node refers to its requirements when making a connection. Our proposed model, the Settling Node Adaptive Model (SNAM), uses a connection algorithm based on connection standards for nodes and does not use the preferential attachment connection algorithm. Thus, SNAM incorporates both a node's properties and heterogeneous connection standards.

4) Each of the proposed models, IASM_A, IASM_B, and SNAM, was validated via simulation using MATLAB [41]. The success of each proposed model to mimic real-world complex networks is verified by examining the statistical properties of the generated synthetic networks, namely the average path length, clustering coefficient, and degree distribution. The statistical properties of the networks generated via simulation are compared to values reported in the literature [2, 4, 5]. Simulation results show that both IASM and SNAM preserve the small world phenomenon, scale-free degree distribution and high average clustering coefficients of real complex networks.

5) We modify the connection algorithm of both IASM models and the SNAM modelto preserve the statistical property of the emergence of community structure. The nodes are divided into classes using a subset of their attribute values. The connection function, *CF*, has an added class similarity term whose value depends on certain class nodes attributes. Both modified models show dense connections between same community members with fewer connections between different community members.

6) We show through mathematical analysis that the degree distributions for both the IASM and SNAM models follow the power law distribution.

7) We present a case study in which we use a Facebook network dataset to illustrate the potential use of our models in modeling real-world networks. A seed network from the Facebook dataset, together with the whole Facebook attribute matrix, are used to grow

synthetic networks via a MATLAB implementation having the same final size as the Facebook dataset chosen for the case study.

Future research can build on this research to, potentially, develop further contributions. For example, the effect of using eigenvector centrality as the structural property measure on the emergence of community structure can be investigated. Additionally, calculating the life of the generated network in terms of its node-indices, where a node index refers to its order of arrival to the network, Examining fairness of the distribution of the added connections among network-nodes along the life of the generated network can be an idea for future work. While our models deal with abstract attributes, investigating the effect of including the nature of attributes in our network generation models is also an area for future look.

1.3. Dissertation Organization

The remainder of the dissertation is organized as follows.

Chapter 2 presents background and related research. The chapter first defines the distinct features of real-world complex networks. Then, an overview of different existing models for the generation of homogenous complex networks is presented. The last part of Chapter 2 provides details of key existing models for the generation of heterogeneous complex networks.

Chapter 3 presents our problem statement. It then discusses the theoretical foundation of our proposed Integrated Attribute Similarity Model (IASM), including the IASM_A and IASM_B variations, and the triad formation step (TFS) modification. This is followed by a discussion of the theoretical foundations of our Settling Node Adaptive Model (SNAM).

Chapter 4 presents our models, flowcharts describing the algorithms, and parameters for the simulation of the proposed models. Simulation results for each model are reported and are analyzed to assess the success for the models in achieving faithful representations of real-world heterogeneous complex networks.

Chapter 5 presents a mathematical analysis of both IASM and SNAM. This mathematical analysis is performed using mean field theory and rate equations to validate that the networks generated by both models have power law degree distributions.

Chapter 6 presents further validation of the potential of using IASM and SNAM in modeling real-world complex networks via a case study. The case study uses a particular Facebook dataset.

We show that SNAM is more suitable for model this particular network than IASM. The case study provides more in depth understanding of SNAM's connection algorithm.

Chapter 7 concludes this dissertation by summarizing our results and providing conclusions. Chapter 7 also outlines some possible future research areas.

Chapter 2. Background and Prior Research

A complex network is defined as a set of many connected nodes that interact in different ways [1]. Complex networks can be seen in different domains, such as social networks, power grids, and food webs [2]. It has been shown that complex networks exhibit distinctive characteristics regardless of the context in which the network exists [2]. This chapter provides an overview of the common characteristics of complex networks and presents a survey of the different research efforts that attempt to formalize and model the distinctive characteristics and behaviors of complex networks.

2.1. Introduction

Complex networks are comprised of sets of numerous interconnected nodes that interact in different ways. Complex networks represent a wide range of complex systems in nature and society. Complex networks are observed in numerous fields such as the Internet, World Wide Web (WWW), social networks, food webs, and many other domains [2, 3, 4]. Complex networks are large, containing from a thousand to several million or more nodes which are connected by edges. In addition to being large, the structure of complex networks is neither completely regular nor completely random. The structure of a complex network results from the fact that complex systems are self-organizing. As a complex system evolves, interactions, usually represented as edges, among its many constituent units, usually represented as nodes, result in an emergent structure with unforeseen properties. While complex networks do share common characteristics with respect to size, structure, and emergent behavior, there is no single general, precise, and accepted definition of network complexity [1]. Given this lack of an accepted definition, differentiating complex networks from other types of networks is difficult. Network nodes and links can represent different entities and relations depending on the analyzed network. For example, in social networks the users engage with each other for various purposes, including business, entertainment, citation, movies, transport, banking, knowledge sharing, and many other activities. The widespread use of complex networks in different fields has made the study of complex networks and their structure an important research topic.

The study and analysis of data extracted from complex networks has revealed a number of distinct features and behavioral patterns that distinguish these networks. Awareness of these

features can lead to an improved understanding of the network's structure and dynamics. Such knowledge can be utilized in different fields to answer questions such as: How can a social network be a mediator for disease transmission? How can the current WWW structure be used to predict future connections between websites? How can critical nodes or links be identified in power grids? How can one deduce new relationships or reveal potential vulnerabilities in a network? Analysis of complex networks analysis can provide insight into the ties and relations linking nodes and an improved understanding of a network's dynamics. Such analysis can enhance decision making dealing with network management and resource allocation in different applications. These are only some of the motivations that make complex networks an important research topic. Within the field of complex network analysis, researchers focus on three main areas [3]:

1. Network statistical analysis and measurement
2. Network modeling
3. Network behavior prediction

We might have perfect knowledge of the parts constituting the network, but its large-scale structure and dynamics may not be immediately obvious [6]. However, certain statistical properties are common to a large number of these networks [3, 4, 5].

Recent enhancements in computational and storage capabilities have made it feasible to pursue the three areas of study listed above. Researchers are now able to gather and analyze large databases resulting from interactions between different nodes in real-world networks. These developments allow researchers to identify the properties of complex networks. Real-world network datasets are often proprietary and hard to obtain. Thus, researchers often study networks using synthetic datasets generated via mathematical models. Knowledge of the properties of complex networks is essential in modeling these networks. Additionally, altering the parameters in a network model leads to the generation of datasets with different properties. These datasets can be used for thorough exploration and evaluation of network analysis algorithms [3, 4].

The study of complex network draws on concepts from graph theory. In graph theory, a network is represented as a set of vertices joined by edges. An edge implies the existence of a relation between the connected vertices. Networks or graphs are of different types. Graphs can be referred to as bipartite graphs if they contain vertices of two distinct types, with edges running

only between nodes of different types. Another type is referred to as hyper-graphs, representing networks having edges linking more than two vertices. Graphs may be static, having a constant network size, or dynamic, having network size that changes with time [3].

Networks can also be classified as heterogeneous or homogeneous. A heterogeneous network is a network with different types of nodes or vertices, while a homogeneous network is a network that has only one type of node vertex. Edges in a network or a graph can be weighted (each edge is assigned a different weight) or un-weighted. Furthermore, edges can be directed (where edges have a direction associated with them) or undirected [3].

Using graph theory concepts, researchers have modeled the structure of complex networks and investigated how different structures can affect interactions in complex networks. Empirical studies of the statistical properties of complex real-world networks represented as graphs have led to the discovery of several common properties for real-world networks.

Section 2.2 starts by defining the common characteristics of real-world complex networks. It concludes by arguing for the importance of defining mathematical models for complex network to facilitate the understanding of the underlying factors that govern their structures and the current and future behavior of the network-nodes. Section 2.3 provides an overview of the different models proposed to generate homogenous complex-networks. It starts by detailing the most influential complex-networks models suggested by Erdös and Rényi (ER) [4, 5], Watts and Strogatz (WS) [4, 5], and Barabási and Albert (BA) [3]. These proposed models cannot incorporate all four of the common properties of complex networks. Section 2.4 introduces different additional efforts for models for the generation of homogenous complex networks. This section first discusses models that are based on modifying the preferential attachment algorithm (PA) of BA [3]. Models that experiment with algorithms for the attachment of new nodes are then discussed. Models that mainly focus on producing networks that preserve real-world network properties of high clustering coefficients and community structure are presented next. Section 2.5 shifts attention to heterogeneous networks. It summarizes prior research that deals with models for heterogeneous networks, which are networks whose nodes are assigned different properties.

2.2. Common Statistical Properties of Complex Networks

Complex networks represented as graphs have been shown to exhibit several common statistical properties, including degree distribution, average path length, clustering coefficient, and community structure. Recently, it was determined that some real-world networks, such as social networks, also exhibit the emergence of community structure [5]. Of course, there may be other statistical properties that are important when analyzing or describing complex networks. Newman [3] states that these additional measures can differ according to the type of the network and the topic being investigated. Thus, for the purpose of our work, we focus on the three statistical properties listed above and the emergence of community structure. An overview of each of these properties is presented below.

2.2.1. Degree Distribution

The degree of a vertex in a network represents the number of connections that the vertex has. The degree of a vertex *j* in an undirected graph is the total number of edges connected to that vertex and it is expressed as k_j. However, in a directed graph, edges are classified as ending at a vertex or as originating from a vertex. The in-degree of a vertex *j* is the total number of edges ending at vertex *j*, while the out-degree of a vertex *j* is the total number of edges originating from vertex *j*. The in-degree and out-degree of a vertex *j* are expressed as $k_{j\ in}$ and $k_{j\ out}$, respectively. Thus, the total degree of a vertex *j* in a directed graph will be expressed as $k_{j} = k_{j\ in} + k_{j\ out}$.

P(k) represents the fraction of vertices in the network with degree *k* and it denotes the probability that if a vertex *v* is picked uniformly at random it will have degree *k*. Degree distributions in complex networks can follow an exponential, Poisson or power law distribution according to the network's nature [3].

2.2.2. Average Path Length

The path length between a pair of vertices is equal to the number of links or hops that form the path that connects the two vertices [3]. There may be different paths connecting a pair of vertices. The shortest path, referred to as geodesic distance, is the connecting path that has the smallest number of links. The average path length in a network is defined as the average number of links along the shortest paths for all possible connected pairs of vertices in the network [3].

For an undirected network having n vertices, the average path length l is the mean of the geodesic (the shortest) distance between all vertex pairs in the network and it is defined as:

$$l = \frac{1}{½n(n-1)} \sum_{i \geq j} d_{ij}, \tag{2.1}$$

where, d_{ij} is the shortest path (or geodesic distance) between any two vertices i and j [3].

2.2.3. Clustering Coefficient

A node's clustering coefficient is defined as "the average fraction of pairs of neighbors of a node that are also neighbors of each other" [2]. Generally, the clustering coefficient is used to assess transitivity of real-world networks. Transitivity means that if vertex i is connected to vertex j, and vertex j is connected to vertex k, then there is a high probability that vertex i will also be connected to vertex k. The value of the average network clustering coefficient, C, ranges between 0 and 1, and can be defined in any of the following ways [3]:

1) $$C = 3 \times \frac{\text{number of triangles in the network}}{\text{number of connected triples of vertices}}, \tag{2.2}$$

   where a triangle contains three interconnected vertices and a connected triple is a single vertex with its two edges running to an unordered pair of vertices[3].

2) $$C = 6 \times \frac{\text{number of triangles in the network}}{\text{number of directed paths of length 2}}, \tag{2.3}$$

   where triangles are as defined above and a directed path of length 2 refers to a directed path of length 2 starting from a specified vertex [3].

3) Watts and Strogatz [4, 5] calculated the network's average clustering coefficient as the average of the individual clustering coefficients of network vertices C_i's. The clustering coefficient for node i is given by:

   $$\text{Node } i \text{ clustering coefficient} = C_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered on } i}, \tag{2.4}$$

   where a triangle has one of its vertices at node i, while a triple is a two-sided incomplete triangle with its vertex at i.

2.2.4. Community Structure

A community is a group of vertices having high density of edges within the group (the community) and a lower density of edges to vertices of other groups (other communities) [5].

Some networks show the presence of communities or a “community structure.” This can be accurately evaluated by using community identification techniques. Community identification is a well-established field of research and an extensive survey is presented in [6]. Networks having a community structure are sometimes referred to as networks with high clustering coefficients. However, according to present definitions, the two properties are not considered equivalent [7].

2.2.5. Statistical Properties of Real-World Networks

As previously mentioned, complex networks are found in many fields. However, there are some real-world complex networks which are frequently cited by the research community [3, 4, 5]. The Internet, social networks, and the World Wide Web represent just a few of the many examples investigated in prior research. We review the empirical properties reported in [3, 4, 5] for the complex networks most frequently cited in the research community. These statistical results are for networks that span several disciplines. Again, the analysis of the statistical properties focused on three statistical properties: average path length, clustering coefficient, and degree distribution [3, 4, 5].

1) World Wide Web (WWW): The World Wide Web is one of the most studied complex systems. The network nodes represent the web pages which are connected by hyperlinks (URLs) that point from one webpage to another in the WWW. The WWW is represented by a directed network with two degree distributions, in-degree and out-degree, which are found to follow a power law distribution. The WWW displays the small-world property in that its average path length scales logarithmically with the evolving network size. The clustering coefficient has been calculated for the undirected version of the network and is found to be higher than that of a random network of the same size [3, 4].

2) Internet: The Internet is an undirected network of physical connections between computers and routers. The structure of the network changes with computers and routers arriving to the network and departing from it. The Internet has higher clustering coefficient values than a random network of the same size. Its average path length preserves the small world property. The degree distribution follows a power law distribution [3, 4].

3) Movie actor collaboration network: In the network representing collaborations of movie actors, actors are represented as nodes and two actors are connected if they have acted in

a movie together. The degree distribution follows a power law distribution. Both the small world property and a high average clustering coefficient are found in this network [3, 4].

4) Science collaboration network: Similar to the collaboration of movie actors, in the science collaboration network, two scientists are connected if they have written an article together. Science collaboration networks have been studied for many scientific fields, including physics, computer science, and more. Networks for all fields studied were found to have small average path length and a high clustering coefficient. The degree distribution is found to be a power law with different exponent values for science collaboration networks in different fields [3, 4].

5) Web of human sexual contacts: This web is typically used to trace sexually transmitted diseases. It has been found to follow a power law degree distribution with different exponent values according to the constructed network [3, 4].

6) Cellular networks: These are directed networks of substrates, such as Adenosine triphosphate (ATP), Adenosine diphosphate (ADP), and water (H_2O), being connected by the chemical reactions in which these substrates can participate. The in-degree and out-degree distributions follow power laws with varying exponents. The undirected version has a small average path length and a large clustering coefficient [4]. The network of protein-protein interactions has a power law degree distribution with an exponential cutoff [4].

7) Ecological networks or food webs: These are directed networks which describe different species connected by edges representing predator-prey relationships. Food webs have high clustering coefficients. Some food webs have a power law degree distribution and for some the degree distributions are exponential [3, 4].

8) Phone call network: The nodes are phone numbers connected by edges or phone calls directed from the calling to the receiving subscribers. Both in-degree and out-degree are found to follow a power law distribution [3, 4].

9) Citation networks: Here, the nodes are different articles and a directed edge represents a citation from a more recent publication to an older related published article. While, the

in-degree distribution follows a power law, the out-degree distribution is exponential for some networks and a power law for others [3, 4].

10) Networks in linguistics: Nodes represent words that are connected when they appear next to or one-word apart from each other in sentences. The network has a small average path length, a high clustering coefficient, and a two-regime power-law degree distribution. Another network linking words with the same meaning (synonyms) has a small average path length, a rather high clustering coefficient, and the degree distribution follows a power law [4].

11) Power and neural networks: The power grid network has generators, transformers, and substations representing nodes which are connected via high-voltage transmission lines. The neural network has neurons as nodes which are linked and two neurons are connected by either a synapse or a gap junction. Both power networks and neural networks have small average path lengths and high clustering coefficients. The power grid has an exponential degree distribution and the degree distribution in neural networks peaks at an intermediate value and then follows an exponential distribution [3, 4].

12) Protein folding: The nodes are the different conformations occurring during folding a protein and two conformations are connected if it is possible to obtain them from each other. The networks studied have the small-world property, high clustering coefficient values, and a degree distribution that is a Gaussian distribution [4].

13) Online social networks: In online social networks, nodes represent people who are connected together by the existence of a relationship between them, such as friendships, business relationships, etc. Social networks have a power law degree distribution with high clustering coefficients and small average path length values [3, 5].

Other examples of complex networks that are less frequently cited according to [3] include networks of company directors, networks of email communications, preference networks (with two kinds of vertices representing individuals and the objects of their preference), networks of airline routes, networks of roads, railways and pedestrian traffic, and the genetic regulatory network.

Observation of the above real-world networks shows the existence of several common statistical properties that differentiate real-world networks from other networks. These properties are: (i) the small world effect, (ii) high clustering coefficient, (iii) scale-free power law degree distributions, and (iv) emergence of community structure. The small world effect means that for a certain fixed value of the mean degree, the value of average path length scales logarithmically, or slower, with network size. The average clustering coefficients in real complex network tend to have high values. In addition it has been observed that real-world networks show an emergent community structure [4, 5]. Furthermore, graphs representing real-world networks have a scale-free power law degree distributions, $P(k) \sim k^{-\gamma}$, where the power law exponent γ is independent of the size of the network and its value is in the range of $1 < \gamma < \infty$ [1,4].

2.2.6. The Need for Mathematical Models of Complex Networks

There is an important need for devising a mathematical model that facilitates performing mathematical analysis on complex networks [2]. Such mathematical models can be used to observe and/or predict how the complex network behaves under different scenarios. Mathematical models can also be used when real datasets are impossible or expensive to gather to generate synthetic datasets that may be used for network analysis. A good mathematical model should successfully mimic the modeled network's statistical properties. Several mathematical models have been proposed to mimic complex networks. The proposed models have been assessed to find out which of the observed real-world network characteristics are incorporated into each model. The validation method for any of these models is based on performing simulations and statistical analysis of the networks generated based on the model and/or using theoretical approaches, such as continuum theory, the master-equation approach or the rate equation approach [5].

The next section summarizes the three most influential models that have been presented for complex networks.

2.3. Most Influential Models for Complex Networks

Efforts to faithfully model complex networks have sprouted several models. The most influential models in the complex-network modeling field are: Erdös and Rényi (ER) [4, 5], Watts and Strogatz (WS) [4, 5] along with its modified version in the Newman and Watts model [4], and

Barabási and Albert (BA) [3]. This section reviews the Erdös and Rényi model, Watts and Strogatz model, Newman and Watts model, and Barabási and Albert model as they are the most widely considered models.

2.3.1. Erdös and Rényi (ER) Random Graph Model

Erdös and Rényi (ER) aimed to study the probable structure of a random graph which led them to introduce their random graph model [3]. They focused on determining the properties of such random graphs using probabilistic arguments. Their model starts with N nodes so the maximum number of connections that can be established is $M = N(N\text{-}1)/2$. Undirected edges are placed between any pair of vertices at random with probability p. The probability of a vertex having degree k in this model, is given by the Binomial distribution, approximated as a Poisson distribution as

$$P(k) = \binom{N}{k} p^{k}(1\text{-}p)^{N\text{-}k} \cong \frac{z^{k}e^{-z}}{K!}, \tag{2.5}$$

where z is the mean of the degree distribution and $z = p\ (N - 1)$. When N tends to ∞, z tends to a constant. Thus, the degree distribution of this model follows a Poisson distribution. The clustering coefficient of this model, $C = z/N$, tends to 0 as N tends to ∞. The average path is given by $\ell = \log N/\log z$, which satisfies the small-world property.

Researchers have used huge databases gathered from real-world complex networks to assess the ability of the ER model to represent complex real networks. Such empirical studies show that the ER model is insufficient to model all four characteristics of real-world networks [4, 5]. Unlike real-world networks, the ER network generated model has a Poisson degree distribution and is characterized by having low clustering coefficients [4, 5].

2.3.2. Watts and Strogatz (WS) Small World Model

Watts and Strogatz noticed that many real-world networks are neither completely regular lattices nor random graphs [4, 5]. Hence, the WS model introduces the idea of rewiring regular lattices to achieve both the high clustering coefficient of regular lattices and to preserve the small world phenomenon of random graphs. The WS model is also referred to as the small-world network model. The model starts with nodes arranged in a regular lattice, where a node is connected to its k neighbors ($k/2$ on either side). The next step of the model is randomly rewiring each edge of the

lattice with probability p, excluding self-connections and duplicate edges. The value of p varies the model between a regular lattice (at value $p = 0$) and a random graph (at value $p = 1$). The average path length and the clustering coefficients of the generated network depend on the value of p. Over a broad interval of p values, the model is characterized by having a short average path length (as small as the random graph's path length) and a much higher clustering coefficient than that of a random graph. Unfortunately, it does not model or represent the scale-free property for a networks' degree distribution. Also, some vertices may become disconnected from the rest of the network upon rewiring [4, 5].

Newman and Watts suggested a slight modification to Watts and Strogatz model to ensure that no vertices ever become disconnected from the rest of the network upon rewiring [4]. In their model no edges are rewired, but shortcut edges are added randomly between vertex pairs chosen uniformly at random. Newman and Watts allow self-connections and duplicate edges, where two nodes can be connected by multiple edges, which can represent the situation where different types of edges connect two nodes. However, the Newman and Watts model also does not generate a scale-free degree distribution network.

2.3.3. Barabási and Albert (BA) Scale-Free Model

The scale-free power-law degree distribution of real complex networks was not evident in the ER or the WS models, rendering both models to be inaccurate in modeling the four characteristics of real complex-networks. This motivated Barbási and Albert to induce the scale-free property for node-degree distribution in their model [3]. Analysis of large databases from real-world complex networks suggests that the degree distribution in these networks decays as a power law (PL). Barbási and Albert's model is motivated by the desire to obtain a model of real-world networks that preserves the power law scale-free degree distribution. Different from the ER and WS models, where the probability of finding nodes with high degree decreases exponentially with degree k, the BA model shows a power law tail. It was shown by Barbási and Albert that to obtain the scale-free degree distribution, the model must possess two properties: (i) growth and (ii) preferential attachment (PA).

Growth reflects the fact that real-world networks are dynamic and nodes are continuously being added to them. Preferential attachment reflects the belief that nodes usually tend to connect to higher-degree, structurally-popular, nodes. Thus, the probability that a new node connects to

preexisting nodes is not uniform and depends on the degree of the preexisting network nodes. Barbási and Albert showed that using either the growth property or PA alone in the model will not generate a model with a scale-free power law degree distribution [3].

The BA model starts with a small number of nodes (m_o), referred to as the seed network. A new node is added at each time step having m links to connect to m old preexisting nodes ($m \leq m_o$). Those m nodes to be connected to the new node are preferentially chosen based on their normalized degree. Thus, the new node connects to an old node i having degree k_i with an attachment probability $\Pi(k_i)$, where $\Pi(k_i)$ is expressed as

$$\Pi(k_i) = \frac{k_i}{\sum_j k_i} \tag{2.6}$$

Networks generated using the BA model demonstrate scale-free degree distribution $P(k) \sim k^{-\gamma}$, known as the "power law." The average path length of the model increases approximately logarithmically with N, thus it exhibits the small world phenomenon. Even though, the BA model has the power law degree distribution, it fails to demonstrate some of the properties of real-world networks shown by empirical results [4, 5]. The PL degree distribution of BA has a constant exponent value $\gamma = 3$, which is different from some real-world networks, where γ ranges from 1 to ∞. The clustering coefficient of networks developed according to the BA model is about five times higher than that of the random graph of the same size. However, it decreases with network size, unlike the small world model, where C is independent of the network size. Furthermore, the model does not show the high clustering observed in real-world networks [3, 4, 5].

The ER, WS, and BA network models [4, 5] all fail to incorporate the emergence of the community structure property. Thus, none of the three models were successful in representing all characteristics observed in real-world networks, namely, the small-world phenomenon, scale-free degree distribution, a high clustering coefficient, and the emergence of community structure [7]. Several models were introduced to remedy the shortcomings of the previous three models, as well as to propose new network evolution algorithms. When evaluating these proposed models, the main focus was whether or not they succeed in incorporating the four properties of real-world networks [5]. Since the BA model, using preferential attachment (PA), was the only model that demonstrated a power law degree distribution, most researchers have adopted the BA model as their starting point and tried to modify the PA scheme used in BA model.

2.4. Homogeneous Complex Network Models Variants

Several variations of the models described above have been proposed in prior work. Section 2.4.1 briefly summarizes some modifications to the PA scheme used in the BA model, whose objective are to obtain a realistic model of real-world complex networks. Section 2.4.2 then discusses some efforts that tried to model complex networks without implementing the PA algorithm. Section 2.4.3 summarizes models which focus on obtaining a high clustering coefficient. The few efforts that have tried to produce models that preserve the community structure property are briefly discussed in Section 2.4.4.

2.4.1. PA Based Models

Although the BA model [4] was able to capture two important features of complex networks, the model was not a faithful representation of real-world complex networks. This motivated several researchers to further explore the nature of BA's preferential attachment function and introduce changes to the BA model. These changes include incorporating some real-world phenomena into the BA model, such as copying links or adjusting accelerated growth of the PA function in the BA model or changing the PA function all together. The following is an overview of these different approaches.

In later work, Albert and Barabási [8] were driven by the fact that the original BA model has the power law distribution exponent γ fixed at 3, while in real-world systems γ is between 1 and ∞ [3, 4]. They raised the question of the BA's model capability to model different types of networks. Their proposed network evolution model includes local events, such as the addition of new nodes and new links or the rewiring of links. They proved that by modifying the occurrence frequency of these local events, their model can generate either a generalized power law or an exponential degree distribution. Their model depends on two probabilities, p and q. Probability p represents the probability by which m new links are added to random nodes using preferential attachment. Probability q represents the probability by which a random link from a random node i is rewired to a preferentially selected node. Finally, with probability $(1 - p - q)$, a new node is added with m preferentially attached links. Using the continuum theory, they proved that networks generated according to their model have a degree distribution that changes from a power law to an exponential distribution depending on the value of m. Their model's exponent γ

in the power law region was found to depend on p, q, and m as: $\gamma = \frac{2m(1-q)+1-p-q}{m} + 1$. Hence, the model's exponent γ ranges between 2 and ∞. Their proposed model shows that local processes can lead to more realistic generated networks. Other statistical properties were not presented in [8].

It was shown via simulation only by Albert and Barabási [4] that networks depending on preferential attachment for evolution result in an exponential degree distribution. Samalam and Vijay [9] concentrated on proving mathematically that preferential attachment PA alone is insufficient to produce scale-free networks. Samalam and Vijay showed that networks with fixed static node size that evolve using preferential attachment show initially scale-free degree distribution that saturates when the network becomes fully connected. Instead of the network growing by adding nodes, it grows in the model proposed by Samalam and Vijay in [9] by adding edges between a fixed number of nodes. At each time step, both ends of a link are connected to two nodes according to nonlinear preferential attachment probability function $P(k)$ proportional to the node degree k. $P(k)$ is given by $P(k) \sim (a + kw)$, where a is a positive constant, typically set to around 1 to prevent node isolation, and w is a positive constant between 0 and 1. Their model allows self and duplicate edges and the authors claim that this does not affect the model's results [9]. The scale-free degree distribution did not appear in this model. Instead, it was shown, using a master equation for the degree distribution evolution that, for $w = 1$, the degree distribution is exponential [9]. Related to [9], is the work by Krapivsky, *et al.* [10] which shows that using a nonlinear preferential attachment function destroys the scale-free nature of the network. They show that evolved networks are scale-free only when the preferential attachment follows an asymptotically linear equation as suggested by BA [4]. It is shown in [8, 9, 10] that this PA, together with the growth of the network size, are essential to generate a network that has a scale-free power law degree distribution.

M.-Y. Wang, *et al.* [11] introduced the idea that connections in citation networks depend on time. Two nodes or papers can have approximately the same degree, but different numbers of citations during each time period. M.-Y. Wang, *et al.* [11] propose a scale-free model for such citation networks, introducing a short-term preferential attachment mechanism (SMPAM), where the preferential attachment function for the studied citation network takes only the recent one-year period into consideration to affect node connections. The model uses the classic BA algorithm,

but the preferential attachment equations consider only the degrees of the nodes in the most recent year. The work uses the mean field theory to calculate the in-degree distribution for the directed citation network. The model generates networks with scale-free in-degree distributions with PL exponent $\gamma = 2$. However, M.-Y. Wang, *et al.* [11] did not introduce a temporal scale-free model. Instead, they depend on snap shots from the network evolution to prove that the power law degree distribution is still preserved [11].

J. Wang, *et al.* [12] provide a model driven by the desire to preserve the three striking statistical characteristics of real-world networks: (i) scale-free, (ii) high clustering coefficient, and (iii) small average path length (APL). Instead of models that calculate the preferential attachment function based on the network's global information, J. Wang, *et al.* [12] have their PA based on fixed-size local network information. J. Wang, *et al.* [12] claim that for the model to be similar to real-world networks, the local world of each node should have a fixed size. They argue that choosing the local world randomly of a fixed size produces a clustering coefficient that approaches zero when the network size is large. The set of nodes with a distance to node *i* that is smaller than or equal to *s* is referred to as the step *s* local world of node *i*. Triad formation is used to preserve the high clustering coefficient of real-world networks. The model starts with an initial ring-shaped seed network. A node arriving at each time step will connect preferentially to the highest-degree node that is part of the local world of a randomly chosen node. The new node then connects with a probability *p* to the highest-degree neighbor of the node to which it has previously attached. Simulation results illustrated that the generated network characteristics depend on the values of parameters *s* and *p* [12]. It was shown that the local world size *s* affects the resultant degree distribution of the model. For example, $s = 1.0$ produces a power law degree distribution. As *s* increases, there is a deviation from power-law behavior. Only the value $s = 1$ tends to increase the generated network's clustering coefficient. The parameter *p* controls the network clustering by allowing the formation of triads. Clustering is found to increase as *p* increases. Simulation results show that the average path length is less than ln *N* where *N* is the network size, so the small world property is preserved [12].

Barbási, *et al.* [13] studied the co-authorship network which, though not different from other real-world networks, has the advantage that connections have explicit time stamps. The studied co-authorship datasets revealed some of the co-authorship network properties as having a power law degree distribution with two different values for the scaling exponents. Co-authorship

network data also exhibits the small world phenomenon. The most important observation derived from the measurements done in the co-authorship network is that its average degree, unlike as suggested by previous models, is not constant. Rather, it increases linearly with time. Barbási, *et al.* [13] reflect the linear increase of average degree with time in the proposed model by using the acceleration of growth mechanism. Accelerated growth refers to adding additional connections to the network besides the ones added initially by the new node. This causes the number of connections to increase at higher rates. The model starts off using the classic BA model, with a new node preferentially establishing *m* connections with preexisting network nodes having the highest degree. Next, internal links are formed among pairs of network nodes according to a preferential function depending on the product of the corresponding degrees of each node pair, the number of newly created internal links per node in unit time *a*, and on the current number of nodes in the network. Using continuum theory, it was shown that the model preserves power law degree distribution and exhibits two scaling regions with exponent $\gamma = 2$ for small node-degree, *k*, values and exponent $\gamma = 3$ for large value of *k*. Both average path length and average clustering coefficient could not be proven analytically and were examined by simulation [13]. Simulation results showed that the generated network diameter is proportional to the logarithm of the network size, and that the clustering coefficient increases as the value of the parameter *a* increases, achieving high clustering coefficients.

Dorogovtsev and Mendes [14] argued that real networks have connections that disappear and connections that are added with time between old network nodes. This phenomenon is referred to as local change. Though the BA model successfully represents growing scale-free networks, it does not represent the decaying (links disappearing) characteristic or the continuous developing (links added between old nodes) nature of real-world networks. The model starts as in classic BA with a node added to the network each time step and connected to an old node with a probability proportional to the old node's degree. The model mimics the nature of developing real networks by allowing *c* new edges to be introduced at every time step. Network growth or decay is indicated by the sign of *c*. For $c > 0$, edges connect pairs of nodes with a probability proportional to the product of the pair of nodes' degrees. Decay of the network structure corresponds to $c < 0$. In the decaying network at every time step $|c|$ edges are removed randomly. Only one action of the addition or removal of links can happen at a time. For $c = 0$, the model is reduced to the basic BA model [3]. The time of birth or arrival of each node *s* is used for labeling the nodes.

Dorogovtsev and Mendes [14] analytically and by simulation show that for developing networks, the degree distribution is a power law with exponent depending on the value of *c*. The decaying structure shows power law scaling only at values of *c* close to zero, only for slowly decaying networks.

2.4.2. Variants to the PA Model

This section introduces models that are able to generate networks having power law degree distributions without including a preferential attachment (PA) algorithm in the model.

Kleinberg, *et al.* [15], studying a directed web graph network, argued that the measurements of the local structure of the web graph suggests that the in-degree and out-degree distributions of the web graph should follow power law distributions. Additionally, a copying mechanism where the new node copies its edges from a random node is essential for the web's content-creation. They proposed a model for growing networks based on the addition of both nodes and edges. The edges are added randomly or using a copying mechanism. The copying mechanism entails randomly choosing a node which connects *m* links to neighbors of other randomly chosen nodes. The new node creates its *m* connections by randomly choosing nodes with probability β. The new node copies its *m* connections from the connection of a random node with probability (1 - β). Kleinberg, *et al.* [15] argue that the copying mechanism is present in a web graph where the pages covering a certain topic are usually linked by interested users and a new page about the same topic is usually connected to them. The model was found to preserve power law distributions using only heuristics. Kleinberg, *et al.* [15] also argued that analytical tools were unable to prove this conclusion because the copying mechanism generated dependencies between random variables. The generated network average path length and clustering coefficients were not assessed.

Krapivsky, *et al.* [16] provide a model that reflects the thought that an author citing a paper is most likely going to cite one of its references as well. Krapivsky, *et al.* [16] proposed a model for re-directed growing networks. In their model, when a new node *i* is added to the network, its edge attaches to a randomly chosen node *j* with probability (1 - *r*). Then, with probability *r*, this edge from the new node *i* is redirected to the ancestor node *o* of the previous randomly chosen node *j*. Node *o* is the ancestor of node *j* if it is the destination of a directed edge from node *j*. The rate equations of the model show that it has a power law degree distribution, with the degree

exponent decreasing with an increase in probability *r*. Other statistical properties were not studied.

BA preferential attachment models use global network information. Global information about the degrees of all nodes in the network is used to calculate all probabilities of linking to them, which seems unrealistic because this can consume a lot of the nodes' resources. Therefore, it is advantageous to propose some network growth models based on local schemes requiring only local knowledge about the vertex under consideration and the nodes closest to it. A model that depends on random walks in establishing connections between nodes only uses such local information. Herrera, *et al.* [17] noted that a network generation model based on a random walk preserves the scale-free degree distribution. The probability of reaching a node *i* of degree k_i in a random walk of arbitrary length *l* is equal to the normalized degree k_i value. Herrera, *et al.* [17] proposed a network generation model that generates a scale-free network with an adjustable clustering coefficient. The value of the clustering coefficient can be increased by increasing the number of triangular connections formed during the random walk. In the random walk based model, each node *x* is assigned probability $P(v_x)$ drawn from a binomial distribution. The new node *i* randomly chooses a preexisting node *j*. Starting from node *j*, a random walk is started of length *l* that ends at node *e*. End node *e* is marked. A new random walk is started from the marked node *e*. This new random walk will be a 1-step walk with probability $P(v_x)$ or a 2-step walk otherwise. The destination node of the random walk is also marked. This process is repeated until there are *m* marked nodes. The new node *i* then connects to the *m* marked nodes. Notice that a 1-step random walk will form a triangle when both of its endpoints are connected to the new node. Thus controlling the number of formed triangles depends on the number of 1-step and the number of 2-step random walks deployed in the model. These numbers in turn depend on each node's assigned probability $P(v_x)$. Clustering control parameter *cc* is defined as the fraction of nodes having $P(v) = 1$,with the remaining nodes having $P(v) = 0$ and the value *cc* is used to control the clustering coefficient. Simulation analysis is presented to validate the model. Simulation shows that the model provides scale-free degree distribution. The model also results in a high clustering coefficient that is proportional to the value of the clustering control parameter *cc* value and is independent of the network size [17].

2.4.3. Enhanced Average Clustering Coefficient Models

Barabási and Albert gave no analytical results for the resulting average clustering coefficient, C, in the BA model, but it is indicated that C decays with the network size N as $C \sim N^{-0.75}$ [17]. However, typical real-world networks have C independent from N and generally higher than the values given by the BA model [4]. Many models attempt to find network models that result in generating networks with high C values as in real-world networks. The following subsections briefly discuss such models.

D. Wang, *et al.* [18] note that in real-world networks not all edges are equivalent and each has its weight. They argue that the weights of the links play a key role in characterizing different real-world networks. D. Wang, *et al.* [18] propose a model to generate scale-free networks that includes the dynamics of weight evolution in its mechanism. The model starts with a fully connected seed network whose links are all assigned a weight of value 1 [18]. The model adds one new node with probability φ, whose value is defined by the model. The new node will be preferentially attached to m old nodes in the original network. The node's strength is defined as the sum of the weights of all links attached to it. The PA probability for each old node is a function of its normalized strength. Also, the model continues to add one new fully connected community of same size m as the seed network to the old network with probability $(1 - \varphi)$. Each added community node is attached via a randomly chosen community node to m preexisting nodes according to a strength driven preferential attachment function [18]. The model of a weighted evolving network of D. Wang, *et al.* [18] was analyzed using the mean field method and continuous time approximation. The model captures power-law distributions of edge strengths and weights as well as node degrees. Although D. Wang, *et al.* [18] state in the introduction of their paper that the model gives high clustering coefficients, no analysis or simulations are given to validate this claim.

The model of Herrera, *et al.* [17], discussed in the previous section, produces scale-free networks whose clustering coefficient can be controlled by an l-step random walk triangle formation using only local information. It is found that connecting the added node to a randomly chosen node then performing an l-step random walk from it to connect the added node to the end node of the random walk generates a triangle for $l = 1$ and no triangles for $l = 2$. Simulation shows that the clustering coefficient linearly increases as the value of the fraction of nodes with a random walk

step of length $l = 1$ value increases. The clustering coefficient is also independent of the final network size, as is seen in real-world networks [5].

Lian-Ren, *et al.* [19] present the Closest Neighbor to Neighbor Strength Driven (CNNSD) model that focuses on representing the dynamic evolution of friendship networks in a social networking site. The model of Lian-Ren, *et al.* [19] incorporates characteristics of friendship networks in which two nodes are connected when they have at least one common neighbor and strength driven attachment exists. In strength driven attachment, new nodes prefer to link to nodes with higher weights and interactions. The model defines three possible states for nodes: disconnected (*d*), potential edge (*p*), or an edge (*e*). A potential edge is an edge that exists between two unconnected nodes with a common neighbor. A node transits from state *x* to state *y* with transition rate $v_{x \to y}$. In the CNNSD model, a new node attaches to a preexisting network node *i* with a preferential attachment function based on the node *i*'s weights with probability (1 - *u*). Additionally, to reflect the fact that nodes with a common neighbor are most likely to link, the new node connects to one of the neighbors of the node *i* with probability *u*. Rate equations for the evolution of the number of nodes with degree *k* and potential degree *k** show that the CNNSD model still preserves PL degree distribution with an exponent that depends on the nodes' transition rates. Simulation results also show that the clustering coefficient of this model decreases with the increase of the seed network size from 500 to 1,000 which is not desirable [19].

Holme and Kim [20] focus in their model on reflecting the high clustering coefficient found in some real-world networks such as social networks. Their model follows the classic BA model with a new added node connecting preferentially to *m* preexisting network nodes. The model introduces a triad formation step (TFS) together with the preferential attachment (PA) step. Triad formation in social networks reflects the tendency of actor *x* after connecting to another actor *y* to connect to the connections of actor *y*. Accordingly, as the new node is added, it is preferentially attached to an old network node *w.* Then, with probability P_t, the new node connects to a randomly chosen neighbor of *w* thus forming a triad (TFS). The new node is linked preferentially to a non-neighbor old network node (PA step) with probability (1 - P_t). This continues until the new node establishes its *m* connections. The model is controlled by the parameter $m_t = (m - 1) P_t$, which is the average number of times that a new added vertex performs a triad formation step.

The clustering coefficient is increased by increasing the number of triads formed. Hence, the parameter m_t controls the value of the evolving network clustering coefficient since it controls the percentage of triads formed. It is shown analytically that the generated network degree distribution follows a power law distribution. It is then shown via simulation that the clustering coefficient increases with an increase of m_t and is independent of the network's size, as in real-world networks [4], and that the small world property is preserved. The emergence of community structure was not discussed by Holme and Kim [20].

Bhukya [21] proposes a model with the objective of including all social networks properties. These properties include high clustering coefficient, small average path length, power law degree distribution and the emergence of community structure. The model reflects when a person supplies someone asking for help with contacts of his or her friends or their friends if he or she was unable to personally provide help. The model has three processes that include a random attachment to an initial contact, an attachment to the neighbor of the initial contact according to PA function (secondary), and, finally, an attachment to the neighbor of the neighbor of the initial contact (tertiary). In Bhukya's Neighbor of Neighbor of Initial Contact (NNIC) model [21], a node chooses, on average, $m_r \geq 1$ random nodes as initial contacts, $m_s \geq 0$ neighbors of each initial contact as secondary contacts, and, finally, an average $m_t \geq 1$ neighbors of each secondary contact as tertiary contacts. The new added node connects to the initial, secondary and tertiary contacts. This process is repeated until reaching the final required network size. Both rate equation analysis and simulation [21] show that the model follows a power law degree distribution with exponent $3+2/m_s$. It is shown that the clustering coefficient of node i depends on its degree k_i as $c_i(k) \sim \ln k_i/k_i$ [21]. Bhukya [21] considers that high clustering indicates the presence of community and that having more than one initial contact node per new node represents a connection between communities. However, the model does not show the emergence of community structure in the sense defined previously [4, 5].

Fu, *et al.* propose a model [22] that uses the Relatively Preferential Attachment (RPA**)** method to generate networks having high clustering coefficient values similar to real networks. Fu, *et al.* [22] note that after the node has made its first connection using PA, its current network location should affect its future connections. Other PA based models neglect the effect of the length of the path that links the old node to the new node. The RPA model gives a higher attachment

probability to an old node that already has one of its neighbors previously attached to the new node. At each time step, the new node connects to *m* old nodes according to the modified preferential attachment probability:

$$\pi(i) = (1-a)\frac{k_i}{\sum_j k_j} + a\frac{h_i}{\sum_v h_v}$$

Here, $0 \leq a \leq 1$ is a parameter that indicates the preference of having larger number of node neighbors h_i over having a high node degreek_i, where h_i is the total number of neighbors, at the time of this link connection, common to both the old node *i* and the new node and $\sum_v h_v$is the sum of h_i of all old preexisting network nodes.

This gives nodes in the immediate neighborhood of the new node a probability of connecting to it higher than other nodes [22]. It is shown using simulation [22] that the power law degree distribution is preserved in the RPA method with exponent greater than 2. The clustering coefficient is found to increase with the increase of the parameter *a* towards real networks values.

Wang and Rong [23] present a model that is based on the tendency of small groups of individuals to first link together before connecting to an existing large complex network. Wang and Rong's evolving small-world networks model is based on a modified BA model [23]. The model starts with a ring seed network having m_o nodes. The basis of this model is that instead of one new added node, a group of *m* fully connected nodes is added at each time step to the network. Each of the *m* nodes in the added group then forms *s* links to the rest of the network nodes following the classic PA rules. Using simulation, Wang and Rong's model [23] shows the scale-free property, high clustering coefficient and small average path length (APL), where the value of *s* is fixed at one. Scale-free PL distribution is maintained for values of $m = 2$ and $m = 3$. But, when $m = 4$, the degree distribution shows a deviation from the scale-free PL distribution. Values of $m > 2$ result in the formation of triads. It is also shown via simulation [23] that, for $m > 2$, the clustering coefficient *C* is relatively stable for different network sizes, *N*. The average path length is shown to be $l < \ln N$. Wang and Rong's model [23] shows the existence of different clique sizes. However, the cliques are spread uniformly over the network and there is no densely connected set of nodes, i.e., there is no community structure.

Jian-Guo, *et al.* [24] modify the Holme and Kim model [20] and introduce the Multistage Random Growing Network (MGRN). While, the attachment to the neighbor in Holme and Kim's model is random, the MGRN model makes attachment to the neighbor following a PA algorithm (MGRN) model that starts with a seed of three nodes. At each time step, a new node is added to the network. This new node connects according to PA function to a preexisting network node and to one of its neighbors chosen preferentially. Analysis uses mean field theory to show that the model's degree distribution is a power law with an exponent $\gamma = 3$ [24]. It is shown that the average path length remain less than ln *N*, where *N* is the network size [24]. The analytically calculated clustering coefficient is equal to 0.83, while simulation results in clustering coefficient equal to 0.74 [24].

Klemm and Eguiluz [25] present a model that attempts to preserve the high average clustering coefficient, small world property, and PL degree distribution found in real-world complex networks. The model divides the network nodes as "active" and "deactivated" nodes during network growth. Upon arrival, the new added node links to *m* "active" nodes in the network. Then the new node becomes "active." The connections in the model are controlled by a connection probability μ. Each of the *m* links of the new node connects with probability $(1 - \mu)$ to an "active" node while it connects with probability μ to a random node ("active" or "deactivated") selected according to a preferential function. The node with lowest degree value is deactivated after connecting all of the *m* links of the new node. The model is examined via simulation [25] which showed it had small average path length, high clustering coefficient, and scale-free degree distribution in the range of $0 < \mu < 1$. Klemm and Eguiluz [25] did not investigate the presence of community structure in their model.

Newman, *et al.* [26] claim that social networks datasets have shortcomings. These shortcomings come from the fact that acquiring social networks datasets usually depends on questionnaires. This can affect the accuracy of the datasets and a tremendous amount of work is required to acquire a medium size dataset. They argue that affiliation network datasets can be used to solve the shortcomings of social networks datasets. However, in affiliation networks, actors are joined together by common membership to groups. Thus, affiliation networks datasets are larger and more accurate datasets of normal social networks. Models of affiliation networks are usually represented via bipartite graphs, having two opposite (or different) types of nodes. Nodes link only to nodes of the opposite type in bipartite graphs. Newman, *et al.* [26] propose a model for

affiliation networks that starts with N unconnected nodes, where N represents the final network size. Each node i is assigned a random number k_i that is drawn from the probability distribution p_k. The probability distribution p_k represents the desired final network degree distribution. Each node i of the N nodes is connected to k_i stubs (ends of edges). Pairs of stubs from the nodes are then chosen and connected. In the affiliation network only nodes of opposite types are connected. The analytical analysis uses a generating function that encapsulates all information of desired p_k. Analysis shows that the model of Newman, *et al.* [26] gives high clustering while preserving the scale-free degree distribution.

Dorogovtsev, Mendes, and Samukhin [27] argue that previous PA algorithms assumed that new nodes have the same properties independent of the current state of the network. They incorporate the idea that nodes added to the network have different random properties depending on the network's state. They support this by pointing out that predecessor inheritance is a feature found in many networks, such as citation and collaboration networks. Thus, each of the old existing nodes inherit some of their predecessors' attractiveness. The predecessor of the node i is represented by the old node that was connected to node i upon its birth. The proposed model has a new node born with a random number of links connected to it. The model has, at the same time, a new link added that connects two old preexisting network nodes. In the model the degree of the newly added node is not constant. Each new node inherits a fraction of the old node's incoming edges by copying them. The value of that fraction follows some probability density distribution that affects the resultant degree distribution. The model captures the high clustering coefficient of real networks.

2.4.4. Enhanced Community Structure Models

Community structure is defined as entailing dense connections between members of the same community and less dense connections between members of different communities [5]. Few papers deal with models that focus on the emergence of real-world community structure.

A software system is defined by Li, *et al.* in [28] as "a system composed of many interacting units (e.g., classes, components and subsystems) and the collaborations among them directly reflect the design, coding, and execution of software." Li, *et al.* [28] are interested in software networks where the nodes denote classes and interfaces in software systems and edges represent dependency relationships between nodes. Li, *et al.* [28] find the statistical characteristics of

software networks to include the small world phenomenon, scale-free degree distribution and modularity. They measure modularity as the fraction inter-community edges minus the expected value of inter-community edges in a randomly connected network with the same community-divisions [28]. In [28], modularity values approaching one indicate a strong community structure and its values for real-world networks range from about 0.3 to 0.7. They argue that the proposed models for complex networks lack modularity which has high values in the directed software networks. Li, *et al.* [28] propose a model for the evolution of software networks whose algorithm was inspired by comparing versions of the software Eclipse. These comparisons show that newly added nodes attach first to modules then attach to the network and that nodes do not attach to the network individually. Thus, the modular attachment model deals with adding groups of nodes (modules) instead of adding individual nodes. The model starts with a seed network and then modules arrive to the network. The constructed module is then attached to the seed network following the BA algorithm. The size of the module starts by one node and keeps increasing with a constant growth rate as the network size increases. When the size of the module becomes more than the seed network size, the module decomposes into a network with a size equal to the seed network's size with the remaining nodes attached to it. The decomposed network is then attached to the seed network following PA rules. Simulation of the algorithm [28] is preformed and compared to data of an actual network in Eclipse. This shows that the modular attachment model had a 0.07 clustering coefficient which, in the studied software network, was acceptable as its clustering coefficient is 0.06. The model has a small average path length. Although the degree distribution of the model is a power law, it is not scale free as the exponent shows dependence on the network size. Modularity is used to assess the division strength of a network into modules. The modularity values of this model are much closer to those in real networks, ranging between 0.54 and 0.57 [28].

Zaidi, *et al.* [29] record some observations about social network structure. Social networks consist of many small densely connected overlapping groups. Within a group, connections are randomly created based on the nodes' interests. The numbers of connections between different groups are much less than connections between nodes in the same group. Zaidi, *et al.* [29] propose a model that incorporates all of these observations. In an effort to generate a community structure, the clique model has cliques of various sizes representing these groups of the real-world. A node can belong to more than one group. This is modeled by merging two nodes from

different groups, so two cliques become joined by a single node. Nodes are assigned connectivity attributes drawn from a power law degree distribution which determines the number of merges for each node. The model chooses two nodes at random and, if they still did not exceed their merge numbers, they are connected. Multiple overlaps can appear between cliques. The model is based on a static network where the final size equals the initial number of nodes. Simulation showed that the model of [29] has high clustering, small average path length and scale free degree distribution. However, Zaidi, *et al.* [29] claim that there is no metric to identify the presence of communities in a network by analyzing the graph on the whole in a global perspective. They use a visual analysis technique that decomposes the topology of the network to show the presence of community structure in the networks generated by the proposed model [29].

2.5. Heterogeneous Complex Networks Generation Models

Most of the proposed models considered homogenous networks, i.e. nodes composing the network are all considered similar. However, there were several attempts that took node heterogeneity into consideration. This section is dedicated to overview heterogeneous complex network generation models. The section discusses models which have nodes with different attributes, attractiveness, age, and capacity.

2.5.1. Node Attractiveness

Dorogovtsev, *et al.* [30] generalize the BA model [3] to account for the different exponent values observed in real networks by associating an attractiveness parameter to the network-nodes. The classic networks evolution model of BA adds one new node at each time step with *m* new links to be linked to old existing nodes. Old nodes are connected based on an attachment probability proportional to their degree. The proposed model [30] assumes that each node is born with an initial constant attractiveness parameter, *A*, to avoid the presence of isolated nodes. Dorogovtsev, *et al.* conclude that the probability that a node receives an incoming edge is proportional to the sum of node's initial attractiveness and the number of its incoming edges. During network evolution, each of the new links is attached to an old node inside the evolving network with an attachment probability proportional to the number of links ending at this old node (the node's in-degree) plus a non-negative initial node attractiveness parameter *A*. This model is equivalent to the BA [3] model if the initial nodes' attractiveness is $A = 0$. This model shows a scale-free

degree distribution with a modified exponent. In Dorogovtsev, *et al.* [30], analysis of the model is accomplished using master equations that are formed and solved to prove that the model has a power law degree distribution exponent of $\gamma = 2 + A/m$. Thus, the power law degree distribution exponent is dependent on nodes' initial attractiveness, A, and the new node's added links, m. The model provides values $\gamma \neq 3$, as observed in some real networks.

In the BA model, as network size increases, the ratio of links to nodes approaches constant m, which is the number of links added with each added new node. Observation of evolving Internet data and WWW data indicate that the ratio of the number of network links to the number of network nodes increases with time. Accelerated growth is used to refer to the number of added links increasing more rapidly than the number of added nodes. In an effort to explain this accelerated growth and its effect on the network's structure, a new model is introduced by Dorogovtsev and Mendes [31]. A new node is attached to m pre-existing nodes via PA as in [30]. Additionally, time-dependent new directed links are distributed among old nodes whose number is time varying as $c_o t^{\alpha}$, where c_o and α are constants whose values control the accelerated growth rate. Each of these added $c_o t^{\alpha}$ links comes out from an arbitrary old node and is directed to another old node s chosen by an attachment probability proportional to $(q_s + A)$, where q_s is the number of in connections to node s (in degree) and A is its initial attractiveness. This accelerated growth model generated a network with a power law degree distribution with an exponent proven analytically to be $\gamma = 1+\frac{1}{1+\alpha}$, which can be different from $\gamma = 3$.

Observations indicate that a node's ability to attract connections does not depend only on degree or age. Nodes in the WWW that provide good content are likely to acquire more connections than others. A new "breakthrough" paper that is part of a citation network is likely to have more connections than an older paper. Thus, each node should be assigned an attribute that describes the competitive nature of the node to make connections. Bianconi and Barabási [32] introduced the term "node fitness." A node i upon birth is assigned fitness factor η_i following some distribution $\rho(\eta)$ that represents its ability to attain connections. The probability of a node with degree k_i to attain connections depends on the product of its degree value and its fitness value η_i. Continuum theory is used to predict the proposed model's degree distribution. Depending on the choice of $\rho(\eta)$, the model can show a power-law degree distribution whose exponent is affected by the fitness distribution $\rho(\eta)$. When $\rho(\eta)$ follows a uniform distribution, the degree distribution

is a generalized power law with an inverse logarithmic correction. Numerical simulations are done to support the predictions of continuum theory [32]. The average clustering coefficient and average path length values of networks generated by this model were not calculated in [32].

Rui, *et al.* [33] introduce node attraction parameter β to reflect a node's capability to attract nodes. Parameter β is defined as the number of connections a node gets in a unit time. The model starts adding new nodes and each node establishes *m* connections. Connections are made according to a PA probability function dependent on the old test node's degree and the node's attraction. Thus, the probability that a new node connects to old test node *i* is given by:

$\pi_i = \frac{\alpha k_i + \gamma \beta_i}{\sum_l (\alpha k_l + \gamma \beta_l)}$, where $(\alpha + \gamma) = 1$.

Mean-field theory shows that having different values of α and γ affects the resultant degree distribution. Numerical simulations show that the model preserved power law degree distribution and has small average path length. The average clustering coefficients of the generated networks are found to decrease with the increase of the network size unlike real-world networks [33].

Cai, *et al.* [34] argue that network's evolution can be affected by factors that delay its growth or damping factors. However, they found that most proposed network generation models focus on factors that accelerate or facilitate a network's growth. Cai, *et al.* [34] focus on these damping factors that can delay network growth or reduce a node's capability to attain connections. They propose an evolving model of online social networks called Damping Factors-based Evolving Model (DFEM). Damping factors are represented by R_1, R_2, and R_3. R_1 is the decline in the initial attractiveness associated with the node upon its arrival. As the heat of the node is defined as the number of connections it makes per unit time, R_2 is the number of zero-heat nodes whose connections are below a specific value *n*. R_3 represents the removal of edges due to irresistible natural factors which Cai, *et al.* [34] assume to be none, so $R_3 = 0$. The model starts like the original BA model with new nodes arriving and establishing *m* connections within the network. The attraction and damping factors, together with the degree of node *i*, are taken into consideration in the PA function used for attaching to node *i* given by

$\pi_i = \frac{\left(k_i + A_i + D_i(t)\right) - \left(R_{1i}(t) + R_{2i}(t) + R_{3i}(t)\right)}{\sum_j \left(k_j + A_j + D_j(t)\right) - \left(R_{1j}(t) + R_{2j}(t) + R_{3j(t)}\right)}$,

where, R_1, R_2, and R_3 are as defined above. A_i is the initial attraction of node i attracting the new node N at time step t_i. $D_i(t)$ is the evolving attraction of node i attracting the new node N from t_i to t. For the proposed online social network new (DFEM) model, analytical and simulation results show that the degree distribution follows a power law whose exponent depends on m and values of attraction and damping factors [34].

2.5.2. Node Age

Amaral, *et al.* [35], suggested that nodes can have aging constraints that limit the addition of new edges to them. Constraints can be related to the node's aging or the cost associated with making new links. Amaral, *et al.* [35] propose a network generation model to test the effect these constraints can have on the generated network degree distribution. The model has two types of nodes, active and inactive. A node becomes inactive and new edges cannot connect to it when it reaches a certain age (aging) or has more than a critical number of edges (capacity constraint). In both cases, numerical simulations [35] indicate that while for small k the degree distribution still follows a power-law, for large k an exponential cutoff develops [35].

Dorogovtsev and Mendes [36] deal with the fact that in reference networks the probability to attach to old references depends on their current degree and age. Thus, they propose a model where connections are made depending on their age as well as on the degree of old nodes. The age dependence is represented in the PA connection-function as $\tau^{-\alpha}$, where τ is the difference between the present time and the node's birth time and $0 \leq \alpha \leq 1$. In their model, each new node makes only one connection to old nodes. They perform analytical analysis and numerical simulations [36]. Results of these analyses show that the model has a power law degree distribution only when $\alpha < 1$ and the degree exponent is a function of α [36].

2.5.3. Node Capacity

Zhang, *et al.* [37] propose a constant capacity restricted BA model (CCRBA) for complex networks. The model evolves from the BA model. Node capacity puts a limit on the node's maximum degree. A node cannot gain connections that exceed its total capacity. The network is also assigned a capacity parameter. The node is allowed to make connections if and only if its capacity, and the whole network's capacity are not exceeded. Zhang, *et al.* [37] study the influence of a node's capacity on the network's evolution and topology since most network

generation models neglect the fact that node capacity can affect the network's topology. Zhang, *et al.* [37] perform numerical simulations for the model. The model has a power-law exponent smaller than that of the BA model. It does not represent some of the characteristics of real-world networks as its clustering coefficient is lower than that of the BA model and its average path length is longer than that of the BA model [37].

2.5.4. Node Attributes

Shaohua, *et al.* [38] observe that nodes with common traits or interests tend to interact. They introduce an evolving model based on attribute similarity between the nodes to study the effect of similarity between nodes on network evolution. Each of the network nodes has an attribute set. Node attributes can be described by a true or false function as in fuzzy logic. True and false functions can be assigned real values in the interval [0, 1]. Shaohua, *et al.* [38] used fuzzy similarity rules to define a similarity function that can be used to assess the similarity between attribute sets of two nodes. A connection is established between nodes if their attributes similarities fall within certain values defined by Shaohua, *et al.* [38]. They use simulation [38] to compare the properties of the network generated by their model to the network generated by the BA model. Despite the fact that this model satisfies the small world property, its degree distribution does not follow a power law [38].

Kim, *et al.* [39] observe that nodes within networks have different attributes and that most models lack a way to represent the effect of node attributes and their interactions on the network structure. They propose the Multiplicative Attribute Graph (MAG) model as a class of generative models for networks having nodes with different attributes. The model focuses on how node attributes interact to give rise to the observed network structure. MAG combines categorical node attributes with their affinities to compute the probability of a connection. Attributes with positive affinity values reflect the idea that for some attributes nodes are more likely to link with nodes having the same value of these attributes (i.e., homophily). On the other hand, for attributes having negative affinities, people are more likely to link to others having a different value of that attribute (i.e., heterophily). For each of the directed graph nodes, L, they define categorical attributes. Attribute values for each node form its affinity matrix. To compute the probability of node i to form a link to node j, the categorical attribute values of the nodes select their corresponding affinity matrix entries. The probability is then computed by multiplying the

selected entries of affinity matrices. Mathematical analysis is done for a simplified undirected network with binary node attributes. The generated network model properties depend on the values of six parameters: n number of nodes, L number of attributes of each node, μ probability that an attribute takes a value of 1, and $[\alpha\ \beta; \beta\ \gamma]$ attribute-attribute affinity matrix where α, β, and γ are constants specified by the model. Kim, *et al.* [39] show, by simulation, that the generalized version of the model has heavy-tailed (power-law or log-normal) degree distributions, small diameters, and local clustering of the edges. This is not a growing network model and does not use the PA algorithm for node-attachment. Hence, this model does not follow the two characteristics of the BA model. Additionally, the model does not consider the structural properties of the nodes while making connections [39]. Kim, *et al.* [39] claim that the high average clustering coefficients property is found in their model. However, the clustering coefficient values for this model are relatively small (0.1 to 0.02) compared to the clustering coefficient values reported by Kim, *et al.* [39]for real-world networks (0.9 to 0.05) when the node-degree varies from 10^1 to 10^3.

Online social networks are characterized by power-law degree distributions, high clustering, and the presence of community structure. Each node or user in social networks is identified by his or her social identity. Li, *et al.* [40] incorporate social similarity in addition to the PA of the BA model. Social similarity is included to investigate if this will result in the emergence of community structure since social characteristics are essential for formation of social connections. Li, *et al.* [40] define a subgraph to be a community in a weak sense if the sum of all degrees within it is larger than the sum of all degrees toward the rest of the network. Every node is identified with a social identity represented by a vector whose elements represent a distinctive social feature. The social distance between two individuals is identified by a distance function that has a value of one if and only if both nodes have similar attribute values. The model initially follows the basic BA algorithm with a new node added at each time step and then it establishes m connections in the network. The new node connects with probability p to the group closest to its social identity and it connects to the other groups with probability $(1 - p)$. Parameter p is the strength of linking via social similarity. The node to be attached to the new node within a group is chosen following the rule of preferential attachment. The new node, after making its first connection to a node within the group, links to one of its neighbors which is randomly chosen with a certain probability *TFp*, which is the triad formation probability. Linking to neighbors of a

previously attached node is repeated until the new node establishes its m links. Using mean-field equations for the model with each node having an attribute vector of length one, Li, *et al.* [40] show that the generated network follows power-law degree distribution. Modularity Q is used to measure the community structure. To calculate Q for a network divided into k communities, a $k \times k$ symmetric matrix $\boldsymbol{e}$ is defined. Element e_{ij} in matrix $\boldsymbol{e}$ is the fraction of all edges in the network that link nodes in community i to nodes in community j. The sum of the diagonal of this matrix is the fraction of edges in the network that connect nodes in the same community. The sum of row (or column) elements a_i is the fraction of edges that connect to nodes in community i. The modularity Q is given by $Q = \sum_i (e_{ii} - a_i^2)$. Simulation showed that Q values increased with p, which agrees with the idea that the community structure is stronger when the preference to homogeneity increases. The generated network follows a power law degree distribution [40]. Li, *et al.* [40] claim that using triad formation produces high average clustering, but they do not present values to validate the claim and they do not measure the average path length for the generated networks. Additionally, the model does not increase the length of the attribute vector to more than one.

2.6. Discussion

As discussed above, none of the models presented in Sections 2.4.1, 2.4.2, and 2.4.3 show the emergence of community structure in the networks generated by the model. Additionally, the research efforts reported in [8], [11], and [14] have studied only the presence of the PL degree distribution and neglected other statistical properties.

Some models depend on including phenomena such as rewiring [8], link addition and removal [14], and accelerated growth [13] and the effect of adding these phenomena on some of the statistical properties is studied. However, the research reported in [8] and [14] focuses mainly on preserving the PL degree distribution. It is shown in [13] that a network with a high average clustering coefficient, a small world property and PL degree distribution can be generated by controlling the accelerated growth rate.

Research results reported in [4], [9], and [10] all conclude that PA is essential with the growth of the network size to generate a network that has a scale-free power law degree-distribution. However, as discussed in Section 2.4.2, some other connection algorithms do not use PA and are still successful in generating a PL degree distribution.

Section 2.4.2 showed that copying mechanisms from a random node [15] or an ancestor of the node [16] were proposed as an alternative for the PA in making connections. The models were successful in generating networks with PL degree distributions, but the research efforts reported in [15]and [16] do not tackle the other statistical properties of the generated networks [4, 5].

Using local information rather than depending on knowledge of the structure of the whole network for making connections is studied in [12] and [17]. However, random walks are used in [17] rather than choosing the local world of a random node. Both models use triad formation to increase the resultant average clustering coefficients of their generated network.

Triad formation is mainly the method used for enhancing clustering coefficient except for research reported in [25], [26], and [27]. Klemm and Eguiluz [25] use the notion of active and deactivated nodes. The generation model of Newman, *et al.* [26] is based on a static network and the model of Dorogovtsev, Mendes, and Samukhin [27] depends on varying the value of the connections of the new nodes. None of the models presented in Sections 2.4.1, 2.4.2, and 2.4.3 show the emergence of the community structure property of real-world complex networks.

The models proposed in both [28] and [29] are inadequate for representing the four statistical properties of complex networks. While, the clustering coefficient values achieved by [28] are acceptable for the software network that was studied by the authors of [28], it is still low compared to other real-world networks. Additionally, the model described in [28] fails to generate the scale-free property of complex networks. The model in [29] has a high clustering coefficient, small average path length and scale-free degree distribution properties. However, [29] does not define a metric for identifying the presence of communities. The model in [29] is a static network generation model and lacks the notion of growing networks.

The heterogeneous complex network generation models in Sections 2.5.1, 2.5.2, and 2.5.3 are not general in that they apply only to networks whose nodes have attraction, age or capacity properties. Defining general attributes for the network nodes is preferable as it will enable us to generate different types of complex networks. The models proposed in [38], [39], and [40] define general attributes for the network-nodes. The model in [38] does not generate a complex network with a PL scale-free degree distribution. Additionally, both the models of [38] and [39] do not include the structural properties of the network nodes in their connection-algorithms. Furthermore, the model proposed in [39] is not a growing network model unlike real-world

complex networks where nodes are constantly added to the network. The model proposed in [40] is successful in preserving the power law degree distribution, but the paper does not present the measured average path length and clustering coefficients. The model in [40] does not generate complex networks whose nodes are assigned more than one attribute.

Thus, a general heterogeneous complex network generation model is still to be found. This model should preserve the four statistical properties of complex-networks. Additionally, the model should also be capable of reflecting the fact that nodes usually are characterized by more than one attribute.

Chapter 3. IASM and SNAM: Heterogeneous Complex Networks Generation Models

3.1. Introduction

Complex networks have four characterizing features: (i) the small world phenomenon; (ii) scale-free degree distribution; (iii) high average clustering coefficients; and (iv) the emergence of community structure. As shown in Chapter 2, the most influential models for complex networks are the Erdös and Rényi (ER) [4, 5], Watts and Strogatz (WS) [4, 5], and Barabási and Albert (BA) [3] models. The ER and WS models failed to produce a network with power law (PL) degree distribution. The BA model uses a preferential attachment (PA) connection algorithm which reflects the belief that nodes usually prefer to connect to higher-degree structurally-popular nodes [2]. The BA model succeeded in preserving the small world phenomenon of real complex-networks as the WS model did. Even though BA model generates networks with a PL degree distribution, it generates networks with an unrealistic constant PL exponent value of $\gamma = 3$. Additionally, the average clustering coefficient for networks generated using BA is lower than that observed in real complex networks of the same size. Thus, the BA model is still inaccurate in representing all four properties observed in real complex-networks. Also, as mentioned in Chapter2, all three models failed to generate a network having a community structure [3, 5]. Thus, each of these models was able to generate networks that show some, but not all, of the four characteristics of complex networks, thus rendering them inadequate for accurately representing complex real-world networks. Many researchers have tried to introduce models that can remedy the shortcomings of the previous three models.

The purpose of many complex network models, especially earlier models, is to develop a mathematical model that preserves statistical properties of real-world networks. However, many more recent models focus on modeling the assembly, growth, or evolution of the network. The approach of modeling network evolution investigates how certain statistical properties emerge in real-world networks [3]. For example, the BA model investigates the mechanism responsible for the existence of a power law degree distribution. Many other models use another modeling approach which targets capturing the dynamics of an evolving network and allow observation of the statistical properties of these evolved networks. This modeling approach is based on the

principle that if the model correctly captures the dynamics that occur during the network evolution, then it will capture the network topology correctly as well.

We observed the fact that nodes in complex networks differ from each other. Specifically, nodes, or entities, in real-world complex networks have different profiles and characteristics. We argue that nodes having different characteristics influence the density and the pattern of connections within a network.

Thus, our proposed models use the growth mechanism and incorporate the heterogeneity of nodes. This enables investigation of the effect of adding heterogeneity in our models on the properties of the generated network. We believe that adding heterogeneity to network generation models will succeed in generating networks that preserve the statistical properties common to real-world networks, unlike BA, ER, and WS. With our models, we try to generate networks with characteristics that resemble as much as possible the statistical properties common to some of the few real-world networks that have received attention from the research community.

Additionally, including heterogeneity of node properties or connection standards in the connection algorithms of our models makes them more suitable for generating the subset of complex networks that exhibit selective linking. Such networks are said to exhibit assortative mixing or homophily [3]. Thus, our research scope is focused on generating mathematical models for real-world networks that exhibit assortative mixing.

Assortative mixing is defined as a bias in favor of connections between network nodes with similar characteristics [3]. In other words, nodes tend to connect with nodes that are similar to them in some aspects. Assortative mixing is found in online social networks, webs of human sexual contacts, the WWW, the movie actor collaboration network, science collaboration graphs, and citation networks [3]. In online social networks, users tend to connect with users who are similar to them in some way, for example sharing interests or located in the same geographic area. Age, race, cultural similarities, and location are factors in choosing partners in webs of human sexual contacts. Language and subject matter play significant roles in the connections between WWW pages. Networks of collaborations between scientists or actors are affected by their interests, such as their research areas and genres of movies, respectively, location, and, language. The same applies for citation networks where papers tend to frequently cite papers dealing with the same or a very similar research subject.

We expect that assortative mixing has a direct effect on the emergence of community structure, a not so frequently discussed statistical property of complex networks. Communities or groups of vertices that are similar in some way tend to have dense connections among each other and less dense connections with nodes belonging to different communities.

Although some heterogeneous complex network generation models were presented in prior work, these models rely on some assumptions about the exact nature of heterogeneity parameter. Examples of the heterogeneity parameters assigned for network nodes in prior work are attraction, age, and capacity [3, 4]. A few models [7, 8, 9] define general attributes for the network nodes. However, none of these models preserved all four of the defined statistical properties common to most real-world complex networks. Moreover, even the models that have been proposed for heterogeneous complex networks do not integrate the heterogeneity of nodes with other structural properties of the network in the analysis and connection algorithms for generating such networks. Also, many existing models are specific for the generation of certain types of complex networks and, thus, are not general.

Thus, a general model for generating undirected heterogeneous complex networks with characteristics of real-world networks showing assortative mixinghas yet to be found. Such a model should preserve the four statistical properties of complex networks. Additionally, the model should be capable of reflecting the fact that each node possesses many different characteristics. The different characteristics of the node should be represented as multiple attributes per node in the mathematical model.

This work presents two general heterogeneous complex-networks generation models. The models are applicable to different complex networks such as the World Wide Web (WWW), and social networks. We present two concepts of node heterogeneity in these models, which are heterogeneity of node attributes and heterogeneity of node connection-standard. Heterogeneity of node connection standard is defined as the difference in each node’s requirements to make a connection. Differences in the properties or the attributes of network nodes reflect heterogeneous node characteristics.

3.2. Heterogeneous Complex Network Generation Models

As previously concluded, a general model for the generation of heterogeneous complex networks generation exhibiting selective linking is still to be devised. Thus, our goal in this research is to

introduce mathematical models that accurately mimic the structure, dynamics, and evolution of heterogeneous complex networks. The models should be able to reflect the four common statistical properties of complex networks. The proposed models are dynamic growing network models where nodes are added to the network at each time step as in the BA model. To achieve this goal, we turn to the fact that nodes in complex networks are different from each other. Specifically, nodes, or entities, in real complex-networks have different profiles and characteristics. We argue that nodes having different characteristics influence the density and the pattern of connections within a network. The notion of node-attributes is used to highlight the node-distinct characteristics. The attribute set of each node is extracted from the characteristics or profiles of that network node. In our model, attributes are assigned randomly to each node upon its birth in (arrival to) the network.

Accordingly, the network graph G in our research is defined by a three-element set $G = \{V, E, A\}$, where V is the set of nodes in the network, E is the set of edges, and A is the set of attribute vectors defining the profiles/characteristics of all the network nodes. The idea of node attributes has been attempted before, but the models in this work are novel in the following ways.

1) The models present a systematic way of defining attributes by incorporating the attribute set in the graph definition.

2) The proposed models are general and do not make any assumptions about the type of the network with assortative mixing.

3) To the extent of our knowledge, our models are the first to integrate the attribute similarity measure and one of the topological popularity measures in the computation of the connection function, *CF*. *CF* values are used in the connection algorithms used to establish links between each new arriving node and the old network nodes.

4) In contrast to other efforts that considered node attributes, each node in the proposed model is assigned an attribute vector having more than one element. Each element in the attribute vector stands for one of node's attributes and attribute vector element values are assumed to be statistically independent.

5) Our second model, SNAM, introduces another aspect of node heterogeneity which is the nodes connection-standard requirement defined above. This concept of heterogeneity was not previously included in prior network generation models.

6) Through the proper choice of SNAM control parameters, the required values of network statistical characteristics can be achieved.

7) Modifying the function used in the connection algorithm of both models results in the generation of networks showing the presence of community structure.

This chapter presents our two proposed models, IASM and SNAM, in Sections 3.2.1 and 3.2.2, respectively. The theoretical idea of each model is discussed in its respective section.

3.2.1. Integrated Attribute Similarity Models (IASM)

Our Integrated Attribute Similarity Models, IASM_A and IASM_B, are based on the Barabasi-Albert model [3] and preserve the two basic ideas of the BA model, network size growth and making connections based on preferential attachment. The BA model is chosen as the basis for the IASM model because the BA preferential attachment model is the only one among the three influential models that succeeded to generate graphs having a scale-free PL degree distribution as seen from Table 3.1. It also reflects network growth where nodes are constantly being added to the network during its evolution.

Table 3.1. Comparison between the BA, ER, and WS Models

	Barabási-Albert	**Erdös-Rényi and Watts-Strogatz**
Degree distribution	Power Law	Poisson
Number (N) of nodes	Growing	Constant
Connection probability	Preferential attachment	Random and uniform

All of our network models start with an initial seed network having m_o nodes interconnected with a randomly chosen number of edges linking random pairs of nodes. An example of a seed network is shown in Figure 3.1. Starting with the seed network, at each time step a new node is added to the network. Each newly added node has m links that it has to make with m different previously existing nodes in the network, where $m \leq m_o$. These m links are connected to m

existing nodes according to a connection algorithm proposed in the subsequent sections. This process of adding a new node with new edges is repeated until the network reaches its final size of N nodes. The connection algorithms differ in how they use the values of a predefined connection function (CF) to make connections. The connection function values are used to test if the new arriving node will connect to the tested old node. CF depends on the aspects that are important to the arriving node when making a connection. These aspects can be the structural popularity of the tested node or its attribute similarity with the arriving node.

The arriving node can represent a new user arriving to the social network. This user starts making connections with different network-nodes based on his/her own preferences. The user can prefer making connections to structurally popular old users. The structurally popularity of the old users can be measured by its number of connections (degree centrality). Thus, a new user to the social network will be more attracted to make a connection with an old user having the highest degree implying highest number of neighbors (first-degree connections). Making this connection will give the new user the chance to reach many other old users.

Additionally, eigenvector centrality of an old node can be also be used as a measure of its structural popularity. A new user may prefer making connection not only to old nodes having the highest number of direct neighbors but also to nodes having less number of direct neighbors that have many connections to their respective neighbors. Thus, eigenvector centrality considers the first degree connections and the second degree connections in evaluating the structural popularity.

On the other hand, the user may prefer connecting with old users having attributes similar to its own. However, the new user at the same time is still concerned with the structural popularity of these users. For example, new user x can have the same attribute similarity value with old users y and z. However, user x prefers will still prefer connecting to the one of them having more neighbors. The user can prefer connecting to structurally popular users that have the most similar number of attributes to it which corresponds to using a connection function depending on the normalized degree multiplied by the attribute similarity measure. Again, the user may prefer connecting to structurally popular users or the ones that have the most similar number of attributes to it which corresponds to a connection function depending on the normalized degree

added to the attribute similarity measure. Therefore, the definition of the equation *CF* can differ depending on the aspects that concerns the new user will making a connection.

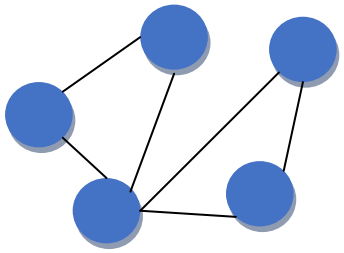


Figure 3.1. Seed network with m_o =5.

In the BA model, each new node is preferentially connected to *m* existing nodes. Each of these connections is based on computing the value of a connection function (*CF*).

The connection function, *CF*, depends only on the normalized degree of the pre-existing tested node in the BA model. The degree of the existing tested is normalized by dividing it by the summation of degree-values of all nodes currently existing in the network. Thus, *CF* is represented as:

CF = Normalized degree of an existing tested node in the growing network

Our first model, the Integrated Attribute Similarity Model (IASM), uses concepts of growth and a PA connection algorithm similar to that in the Barabási-Albert (BA) model for network generation. Our IASM is based on modifying *CF* of the BA model. Instead of having *CF* depend only on the existing node's fitness or degree alone, we propose making *CF* also dependent on a parameter showing the attribute similarity or compatibility between the newly added node and existing nodes in the network. Thus, for each of the required connections of the new node, IASM integrates the attribute-similarity measure between the new node and the existing node with the structural popularity measure of the existing node used to evaluate *CF* used in the PA connection algorithm. The node structural popularity is a measure of the node's popularity based on its network position and connections. To the extent of our knowledge, IASM is the first network model to integrate an attribute-similarity measure within the connection function. IASM has two variations, IASM_A and IASM_B, that use two different structural popularity measures. In IASM_A, the normalized node degree is used as the structural popularity measure. The attribute similarity is calculated by finding the similarity or compatibility between the attributes of the

nodes to be connected. Each of the network-nodes is assigned an attribute vector. This attribute similarity is calculated as the normalized summation of the inner product attribute vectors of the new node and the existing node. Thus, the connection function value for arriving node i and pre-existing node$j$ is expressed as:

$$CF = (\alpha) \times \text{Normalized [(degree of node } j) \times (\text{Attribute Similarity between nodes } i, j)]$$
$$+ (\beta) \times \text{ Normalized (degree of node } j)$$
$$+(w) \times \text{Normalized (Attribute Similarity between nodes } i, j), \quad (3.1)$$

where $\alpha + w + \beta = 1.0$, $0 \leq \alpha \leq 1$, $0 \leq w \leq 1$, and $0 \leq \beta \leq 1$.

The coefficients α, w, and β are weighting coefficients used to define the contribution of the different *CF* equation terms to the final value of *CF* to test their influence. These different terms in *CF* represent the aspects that are considered by the new arriving node when making a connection. These aspects depend on the structural popularity of the tested node and/or its attribute similarity with the arriving node.

Eigenvector centrality is used in IASM_B. Eigenvector centrality is considered a more accurate structural popularity measure as it takes into consideration both the density and quality of links attached to a node. Eigenvector centrality generally assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. More specifically here, a connection to a more interconnected node contributes to the node's Eigenvector centrality to a greater extent than a relationship to a less well interconnected node. A node's eigenvector centrality is calculated by finding the eigenvectors corresponding to the eigenvalues of the network's adjacency matrix. Hence, for IASM_B, the connection function value for a new node *i* and an existing node *j* is measured as:

$$CF = (\alpha) \times [(\text{Eigenvector centrality of node } j) \times (\text{Attribute Similarity between nodes } i, j)]$$
$$+ (\beta) \times (\text{Eigenvector centrality of node } j) + (w) \times \text{Normalized (Attribute Similarity between nodes } i, j) \quad (3.2)$$

To further enhance the clustering coefficient values in the two IASM models, the well-known triad formation step (TFS) has been added to their network generation algorithms. TFS reflects

the preference of a node to connect to its neighbor's neighbor rather than to any other randomly chosen node.

3.2.2 Settling Node Adaptive Model (SNAM)

Our second proposed model departs from the classic PA connection algorithm presented in BA and proposes a new settling node adaptive model, referred to as "SNAM." SNAM reflects the idea that nodes are not only differentiated by their attributes, but also according to their connection-standard requirements. Connection-standard requirements for the nodes represent the minimum *CF* values that a new node finds satisfactory to connect with another old node. The connection-standard requirement of node *x* is its minimum acceptable value of *CF* for establishing a connection. This minimum acceptable value can refer to the value of the popularity measure of a webpage or the value of a similarity measure between a node and other social network nodes.

All of the proposed IASM models assume that all the arriving nodes have the same requirements for the existing old nodes to which they connect. In real life this is not always true as some new arriving nodes can have lower connection requirements to existing nodes than others. Two newly-arriving nodes may have different connection standards, thus one node might accept an obtained *CF* value and make a connection to tested existing node, while the other node might reject the same *CF* value and refuse the connection to the same tested existing node.

To reflect this behavior in our SNAM model, each new arriving node is assigned the value of its own intrinsic connection standard upon birth. This connection standard is used by the node only upon arrival to make a decision about which connection to make with randomly chosen existing old network node. If the *CF* value is equal to or higher than the standard of that arriving node and the arriving node has not yet established its m connections, then the two nodes are connected. If the *CF* value is below the standard of that arriving node, no connection is made. Then the new node must test other old existing nodes to find the ones satisfying its standard. This is repeated for a finite number of tests. After this finite number of tests, the new node must lower its standard if it did not make its *m* connections.

An extension of SNAM was needed to make networks generated using extended SNAM approach the characteristics of the dataset used in the case study that was part of our research.

We will see later in Chapter 6 that this extension for SNAM is introduced by adding a new model parameter *R* to control the number of reductions of the standard of the arriving node.

Thus, an arriving node *x* calculates its connection function value with an existing node *z* (CF_1). If CF_1 is equal to or higher than the arriving node's *x* connection standard, then node *x* makes a connection to node *z*. If another new node *y* calculates its connection function with the same old test node *z* (CF_2) and finds CF_2 lower than its connection standard. Then, node *y* refuses to connect to the same existing node *z*. The *CF* used in our SNAM model depends on the structural popularity of the tested existing node and its attribute similarity with the new node as defined for the IASM_A model, Equation 1 above.

To evaluate our models, we generate networks based on each model using MATLAB [41]. For each of the generated networks, values for the power law exponent, the average path length and the average clustering coefficients are computed to be assessed against values reported for a variety of real complex networks [3, 4]. These statistical properties are the three metrics that validate that the three features of real complex networks are preserved in our models. Our mathematical models are general and apply for any complex network. Upon establishment of the mathematical model, we apply it to social networks as a proof of concept. Our choice of online social network is mainly due to their prevalence and their currently wide application in fields such as marketing, information, diffusion of epidemic diseases, and recommendation and trust analysis.

Chapter 4. Simulation Results and Validation

4.1. Introduction

In this chapter, we present our two proposed models, Integrated Attribute Similarity Model (IASM) and Settling Node Adaptive Model (SNAM). Each of these models is used to generate simulated networks via MATLAB [41]. The statistical properties of the generated networks are recorded. As our goal is to devise a model that mimics the statistical properties and dynamics of complex networks, we calculate values for the power law exponent, the average path length, and the average clustering coefficients. These values are then assessed against values reported for a variety of real complex-networks [3, 4]. IASM is discussed in Section 4.2 along with simulation results and findings. The normalized node degree and eigenvector centrality are used as structural popularity measures in the IASM_A and IASM_B models, respectively. The SNAM model and its simulation results and findings are presented in Section 4.3. Normalized node degree is used as the structural popularity measure in the SNAM model in addition to the connection standard parameter capturing the heterogeneity of nodes. Section 4.4 represents modified versions of both IASM and SNAM that grow networks that exhibit the presence of community structure. Simulation results for such community structure presence is given for both models.

4.2. Integrated Attribute Similarity Models (IASM)

4.2.1. Model Assumptions

The *CF* is defined as the normalized degree only in the BA model. However, attributes are integrated in the *CF* in IASM model to introduce the compatibility parameter that would replace the fitness parameter previously introduced in [32]. This compatibility parameter is a measure of the similarity between the new added node and the old node attributes.

Thus, each new network-node upon birth possesses its own distinct attribute-set (attribute vector) of length L. This attribute vector represents the interests or engagements of that node in the network's L interests or activities. The *CF* in IASM does not depend solely on a specific characteristic of the old or existing node, but on the characteristics of both the new and the existing nodes. Accordingly, a new node usually prefers to connect with existing nodes that are

the most topologically popular and that have similar interests or attributes to the new added node. If there are two existing nodes that possess the same attribute similarity measure value with the new node, then the new node will prefer connecting with the more structurally popular one.

IASM is a growing network model. IASM starts with a seed network of size m_o. Then, at each time step, a new node is added with m edges to be connected to it, where $m \leq m_o$. Each added node is assigned an attribute vector having L elements. Each of these elements takes binary values of 1or 0 representing the presence or absence of an attribute in the attribute vector, respectively. Our proposed compatibility measure represents the similarity between the new added node and an existing test node's attributes. This measure is equal to the normalized summation of the inner product attribute vectors of the new node and the existing node. This attribute similarity measure is integrated within the new defined *CF*, together with structural popularity. Each newly added node is preferentially connected to m old nodes based on the value of function *CF*.

Thus, the connection function *CF* used for preferentially connecting a new node i with a chosen old node j depends on the structural popularity of node j (SP_j) and node attribute similarities (Aij) for both nodes i and j. The connection function *CF* is expressed as:

$$CF = (\alpha) \times \left(\frac{SP_j A_{ij}}{\sum_j SP_j A_{ij}}\right) + (\beta) \times \left(\frac{SP_j}{\sum_j SP_j}\right) + (w) \times \left(\frac{A_{ij}}{\sum_j A_{ij}}\right), \quad (4.1)$$

where $\alpha + w + \beta = 1.0$, $0 \leq \alpha \leq 1$, $0 \leq w \leq 1$, and $0 \leq \beta \leq 1$.

Parameters α, w, and β are weighting coefficients used to give different weights to the combined structural popularity and the attribute similarity in the *CF* terms to test their influence.

The IASM_A model follows the assumption of the BA model where topological node popularity is measured by its degree centrality (normalized degree). For the IASM_B model we argue that the node's topological popularity is better represented by its eigenvector centrality value. Our argument is strengthened by the fact that a node's eigenvector centrality value represents not only the number of that node's connections, but, also, the quality of these connections. Eigenvector centrality as a measure of node's importance is dependent on a nodes own connections and the connections of the nodes connected to that node. A network node has a high eigenvector centrality value if it is connected to many nodes or to a few nodes that have many

connections. Hence, eigenvector centrality seems a better measure of node popularity and is considered a more comprehensive version of degree centrality.

4.2.2. Simulation Setup and Parameters

Simulation of the IASM_A and IASM_B models starts with a seed network of size m_o = 5. The network size grows as new nodes arrive to the network, until reaching a predetermined final size of N nodes. In our simulation N = 1000. Each newly arriving node has to establish m links with the existing network nodes, where $m = m_o$ = 5. Each new node in the network is randomly assigned an attribute vector of length L = 10, whose elements are derived from a uniform distribution. Simulation parameter values used are summarized in Table 4.1.

Table 4.1. Simulation Parameter Values

m_o	m	L	N
5	5	10	1000

The connection function in IASM depends on the attribute similarity between newly arriving nodes and old or existing network nodes as well as the structural popularity of old nodes. *CF* is used to establish node connections preferentially. We use the algorithm proposed by Newman [4] to implement the preferential attachment. Each node is identified by a Node-Id that represents its arrival order. A list of Node-Ids is created for each arriving node in which Node-Ids are repeated based on their corresponding *CF* values. Thus, a new vector is formed in which nodes having higher CF values are repeated more frequently. Each arriving node has to establish m connections with nodes randomly selected from this new vector.

Two different structural popularity measures are used in the simulation of IASM. In IASM_A, a node's structural popularity is based on the node degree. In IASM_B, the structural popularity is based on the node's eigenvector centrality. A flow chart of the algorithm used in both IASM models is shown in Figure 4.1.

MATLAB simulations were performed for different combinations of CF coefficients for both models. Each simulation experiment is repeated 10 times with different seeds for the random value seed generator. The simulation results shown in the tables are present for the average of these 10 experiments. The connection functions used can be based on normalized degree only (β = 1, $\alpha = w$ = 0), on degree with added attribute similarity (α = 0 and w = 1 - β where $0 \leq \beta \leq$

1), and on degree and degree multiplied by the attribute similarity (w = 0, α = 1 - ß, where 0 ≤ β ≤ 1). Parameters α, β, and w are weighting coefficients as discussed in Section 3.2.1.

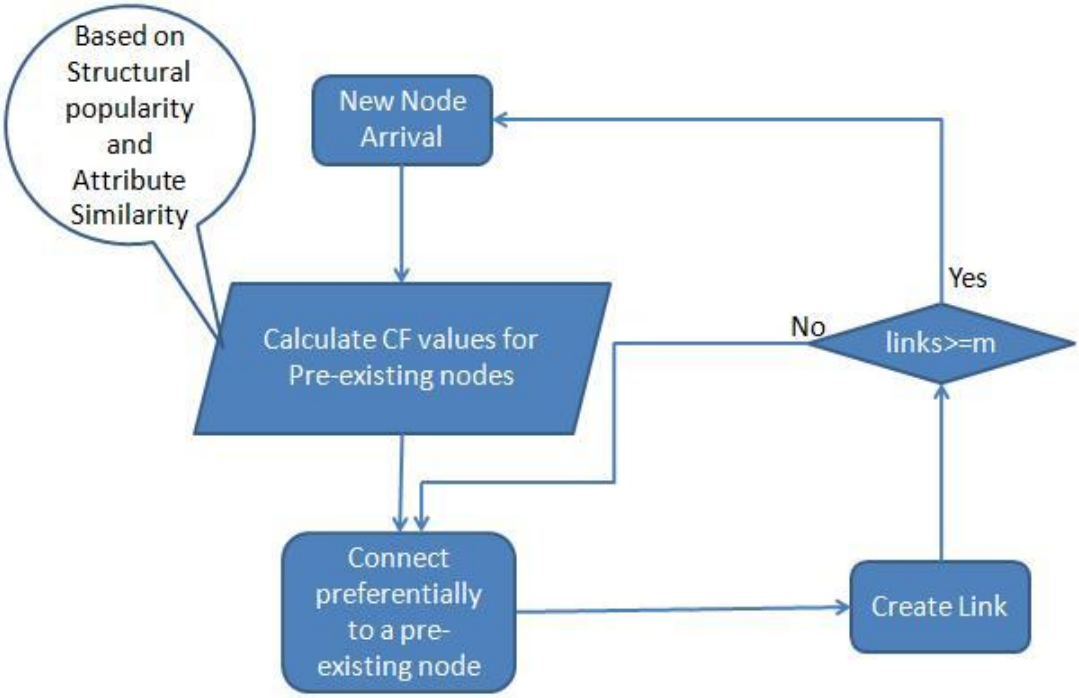


Figure 4.1. IASM_A and IASM_B algorithm flow chart with modified PA function based on: (i) Normalized degree and attribute similarity for the IASM_A model and (ii) Eigenvector centrality and attribute similarity for the IASM_B model.

4.2.3. IASM_A

The following sections introduce the simulation results for the IASM_A model and an analysis of the results that were obtained.

4.2.3.1. Simulation Results

Simulation results for the average clustering coefficient (Av_CC), the average path length (Av_Pl), and the exponent of PL (Exp_PL) corresponding to different values of the weighting coefficient of *CF* are presented below.

4.2.3.2. Analysis of Simulation Results

Model IASM_A reduces to the BA model when β=1. The degree distribution of IASM_A was found to follow a power law distribution whose exponent values are in the range of $2.06 \leq \gamma \leq 2.49$. The average path length values are less than or equal to the logarithmic value of N (1000). Thus, the small world phenomenon is preserved. The average clustering coefficient for the BA model (β = 1) has the value of 0.032. This value increases when the *CF* is based on normalized degree with multiplicative attribute similarity.

Table 4.2. Simulation Results for IASM_A

A	***w***	***B***	**Exp_PL**	**Av_Pl**	**Av_CC**
0	0	1	2.49	3.03	0.032
0.2	0	0.8	2.44	3.02	0.032
0.5	0	0.5	2.39	3.02	0.034
0.8	0	0.2	2.41	2.98	0.041
1	0	0	2.33	2.96	0.044
0	0.5	0.5	2.14	3.14	0.021
0.5	0.5	0	2.06	3.13	0.022

4.2.4. IASM_B

The following sections introduce the simulation results for the IASM_B model and an analysis of the obtained results.

4.2.4.1. Simulation Results

Simulation results for the average clustering coefficient (Av_CC), the average path length (Av_Pl), and the exponent of PL (Exp_PL) corresponding to different values of the weighting coefficient of the *CF* are presented in Table 4.3.

Table 4.3. Simulation Results for IASM_B

α	***w***	***β***	**Exp_PL**	**Av_Pl**	**Av_CC**
0	0	1	2.48	3.04	0.032
0.2	0	0.8	2.43	3.04	0.031
0.5	0	0.5	2.53	3.01	0.031
0.8	0	0.2	2.44	3.04	0.031

1	0	0	2.20	2.97	0.042
0	0.5	0.5	2.02	3.14	0.019
0.5	0.5	0	1.68	3.28	0.014

4.2.4.2. Analysis of Results

The degree distribution of IASM_B follows a power law distribution. The exponent values are slightly less than that of BA model (IASM_A when $\beta = 1$). The average path length values are less than or equal to the logarithmic value of N (1000). Thus, the small world phenomenon is preserved. The average clustering coefficient is still as low as in BA model, but it increases when *CF* is based on eigenvector centrality and multiplicative attribute similarity.

4.2.5. Discussion of Results

The results show that the values recorded for network statistical parameters are similar in IASM_A and IASM_B, which means that the method used in measuring a node's structural popularity has only a minor effect on the statistical parameters of the simulated network. Inducing attribute similarity into *CF* preserved the small world phenomenon, while slightly decreasing the average path length in the case of multiplicative attribute similarity based *CF*. Moreover, the power law exponent values for both IASM_A and IASM_B are within the values reported in [2, 4, 5] for all *CF* coefficient variations in both models. However, incorporating multiplicative attribute similarity in the *CF* calculation had a positive effect on the average clustering coefficient. Simulation results for the average clustering coefficient in IASM_A and IASM_B when (when $\alpha = 1$, $\beta = w = 0$) show a 37 percent and 31 percent increase, respectively, over the BA model. The average clustering coefficient was found to increase with increasing the value of α when $w = 0$ and $\beta = 1 - \alpha$.

However, using additive attribute similarity resulted in a decrease in the generated network's average clustering coefficient values. Thus, we argue that the multiplicative attribute similarity measure is a better measure for similarity than additive attribute similarity in IASM.

4.2.6. Enhancing IASM Clustering Coefficient

The IASM_A and IASM_B model both show low clustering coefficient values. The clustering coefficients can be increased by adding a triad formation step (TFS). The triad formation step is

motivated by the observation that nodes usually form connections with the neighbors of their neighbors.

4.2.6.1. Simulation Setup and Parameters

To form a triad, a newly arriving node attaches to a randomly chosen second-degree neighbor node, and then a neighbor of this existing node is randomly chosen and is also connected to the newly arriving node. Simulation of the IASM_A and IASM_B models is repeated after adding a TFS using the same m_o, m, L, and N values as shown in Table 4.1. The flow chart of the modified model after adding the TFS is shown in Figure 4.2.

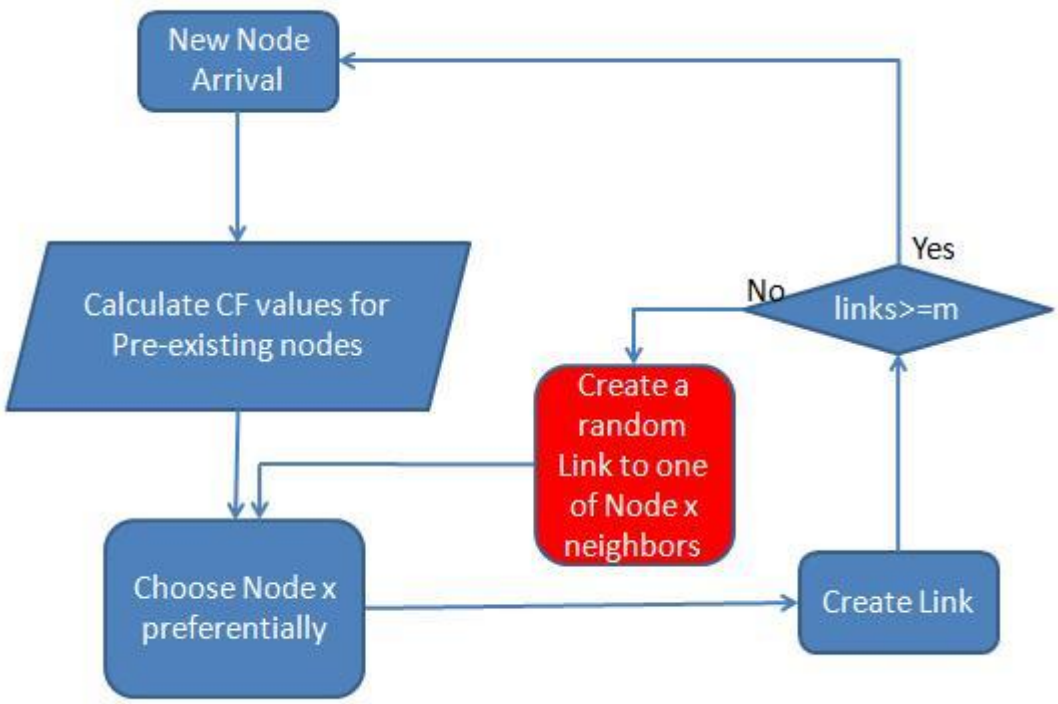


Figure 4.2. Flow chart for modified IASM_A and IASM_B models with triad formation step. CF based on: (i) normalized degree and attribute similarity for the IASM_A model and (ii) eigenvector centrality and attribute similarity for the IASM_B model.

4.2.6.2. Simulation Results

Simulation results for the average clustering coefficient (Av_CC), the average path length (Av_Pl), and the exponent of PL (Exp_PL) for both the IASM_A and IASM_B models after adding the TFS are presented for different values of the weighting coefficient of *CF*.

Table 4-4. Simulation Results of IASM_A and IASM_B Models after Adding TFS

Connection function (*CF*) coefficients			**IASM_A with TFS**			**IASM_B with TFS**		
A	***w***	***β***	**Exp_PL**	**Av_Pl**	**Av_CC**	**Exp_PL**	**Av_Pl**	**Av_CC**
0	0	1	1.91	3.51	0.526	1.93	3.43	0.537
0.2	0	0.8	1.89	3.52	0.526	1.96	3.42	0.535
0.5	0	0.5	1.89	3.52	0.525	1.96	3.4	0.539
0.8	0	0.2	1.90	3.46	0.526	2.00	3.43	0.535
1	0	0	1.88	3.46	0.526	1.97	3.33	0.536
0	0.5	0.5	1.78	3.58	0.515	1.73	3.55	0.518
0.5	0.5	0	1.76	3.58	0.515	1.58	3.68	0.520

4.2.6.3. Analysis of Results

Adding the triad formation step increases the average clustering coefficients values for both the IASM_A and IASM_B models as shown in Table 4.4. The addition of the TFS step to both IASM models generates networks that have power law degree distributions. The TFS causes the PL exponent to decrease to a value below 2unlike both IASM models. The decrease of PL exponent of the degree-distribution values implies an increase in the number of higher degree nodes and thus, the formation of hubs. The reason for the increase in the number of higher degree nodes can be the addition of more connections to the first-degree neighbors of the previously connected node. However, the PL exponent values recorded are similar to some of the real complex networks reported in [2, 4, 5]. In addition, the triad formation step increases average path length while preserving the small world phenomenon. The increase in the average path length could be a result of nodes making more connections with their second-degree neighbors rather than making preferential connections. The addition of the TFS step increases the average clustering coefficient for all combinations of *CF* in both IASM models by nearly 0.5. This suggests that the effect of the addition of the TFS on the average clustering coefficient is almost independent of *CF* coefficient values.

4.3. Settling Node Adaptive Model (SNAM)

4.3.1. Model Assumptions

SNAM introduces the idea of heterogeneous connection-standard requirements of nodes. As previously defined, the connection-standard requirements of nodes represent the different requirements of the nodes when establishing connections. For example, two new users in a social network can have different standards for making a connection. One of these users can accept making connections with only very popular users while the other is satisfied with making connections with less popular users.

To the extent of our knowledge, all previously proposed models assumed that all new arriving nodes have the same connection requirements when linking to existing nodes. In reality, nodes may have different views of the same value of a connection-function (*CF*) that is calculated based on attribute similarity and/or structural popularity of the old node with which to connect. For example, a user in a social network can consider a *CF* value of 0.5 too low, while another user will consider the same value sufficient for establishing a connection with another user. Thus, in SNAM, each arriving node, upon birth is assigned a value representing its own connection standard value *S* derived from a uniform distribution. An arriving node will calculate its *CF* values with existing nodes. Hence, the *CF* values obtained will not be used to deploy the preferential attachment algorithm, but will be used to examine if the randomly chosen existing nodes will meet the arriving node's standards. A newly arriving node calculates the *CF* corresponding to randomly chosen nodes. This characteristic parameter *S*, $0 < S \leq 1$, represents the minimum acceptable value of the *CF* for the new node. All old pre-existing nodes whose *CF* values for the new node are below its standard cannot attach to that new node. An arriving node must then test other pre-existing nodes to find the ones that satisfy its connection standard. The new node establishes connections with the existing nodes whose *CF* values are equal to or higher than its connection-standard value, *S*. Similar to the IASM_A, the *CF* that is used depends on the normalized degree values and/or attribute similarity.

4.3.2. Simulation Setup and Parameters

The network starts with a seed network m_o. A new node arrives at each time step and each new node i is assigned a random connection-standard value S_i, where $0 < S_i \leq 1$. If, for a chosen

existing node *j*, the value of CF_{ij} exceeds or is equal to S_i, then node *i* establishes a connection to *j*. Otherwise, *i* rejects the connection to *j* and another existing old node is tested.

This testing of other existing nodes continues until the new node achieves its predefined *m* connections or reaches its maximum number of tests, *NoT*. If node *i* reaches its maximum number of tests, *NoT*, and it still did not make its *m* connections, then arriving node *i* reduces its connection standard by a certain percentage and the testing of randomly chosen existing nodes is resumed. The reduced standard-connection value, $S_{i,\text{reduced}}$, is determined as follows.

$$S_{i\ \text{reduced}} = S_i \times (1-\epsilon), \text{ where } \epsilon < 1.0 \qquad (4.2)$$

In SNAM, we experiment with the maximum number of tests *NoT* allowed for the arriving node *i* before it has to lower its connection standard if node *i* has not established its *m* connections during the *NoT* tests. As for the previous simulation experiments, the simulation experiment starts with the same values of the parameters α, β and *w*as in Table 4.1. Also, $\epsilon = 0.1$. We choose this value for ϵ because higher ϵ values would make the nodes reduce their standard very rapidly. This will decrease the effect of the presence of the node's connection standard on the generated network. Thus, when ϵ approaches the value of 1, the generated network approaches a network generated with new nodes having no connection standard and this will result in random connections that are not the result of SNAM model. The SNAM model algorithm is shown in the flow chart in Figure 4.3.

NoT is an integer value greater than 1, whose maximum value is the current size of the network. (This implies testing the *CF* values for all network nodes.) Thus, the value of *NoT* depends on the current size of the network, denoted *CS*. The *NoT* is increased by one whenever *CS* of the network reaches certain predefined milestones. The value of the *CS* at which these milestones occur depends on the final size of the network, *N*, and the number of milestones, *NM*, occurring during network evolution. The number of milestones, *NM*, has two extreme values. The smallest number of milestones is 1which is reached when the network reaches its final size. The largest number of milestones occurs when we consider the arrival of each node to the network as a milestone. Thus, *NM* ranges between 1 and *N*. The higher the value of *NM*, the more rapid is the increase in *NoT*.

Our experimentation with the *NoT* parameter indicated that a rapid increase of *NoT* with network growth results in the presence of irregularities in the statistical characteristics of the generated

network. Here, the number of milestones occurring during the arrival of every 100 nodes to the network is varied between 1 and 10. Thus, an *NM* value of 5 means that the *NoT* is increased by one 5 times during the arrival of 100 nodes to the network, (i.e., *NoT* increased by one each time 20 new nodes arrive to the network). This choice was made to avoid irregular statistical properties and has proven to give satisfactory results as shown in Figures 4.4, 4.5, 4.6, and 4.7. The same values for parameters m_o, m, L, and N used for the simulation IASM_A are used for the simulation of SNAM. The initial value of *NoT* in this simulation is 2 and $1 \leq NM \leq 10$. The connection function *CF* depends on the existing node degree structural popularity as in IASM_A, D_i, namely:

$$CF = (\alpha)\times\left(\frac{D_j A_{ij}}{\sum_j D_j A_{ij}}\right) + (\beta)\times\left(\frac{D_j}{\sum_j D_j}\right) + (w)\times\left(\frac{A_{ij}}{\sum_j A_{ij}}\right), \tag{4.3}$$

where $\alpha + w + \beta = 1.0$, $0 \leq \alpha \leq 1$, $0 \leq w \leq 1$, and $0 \leq \beta \leq 1$.

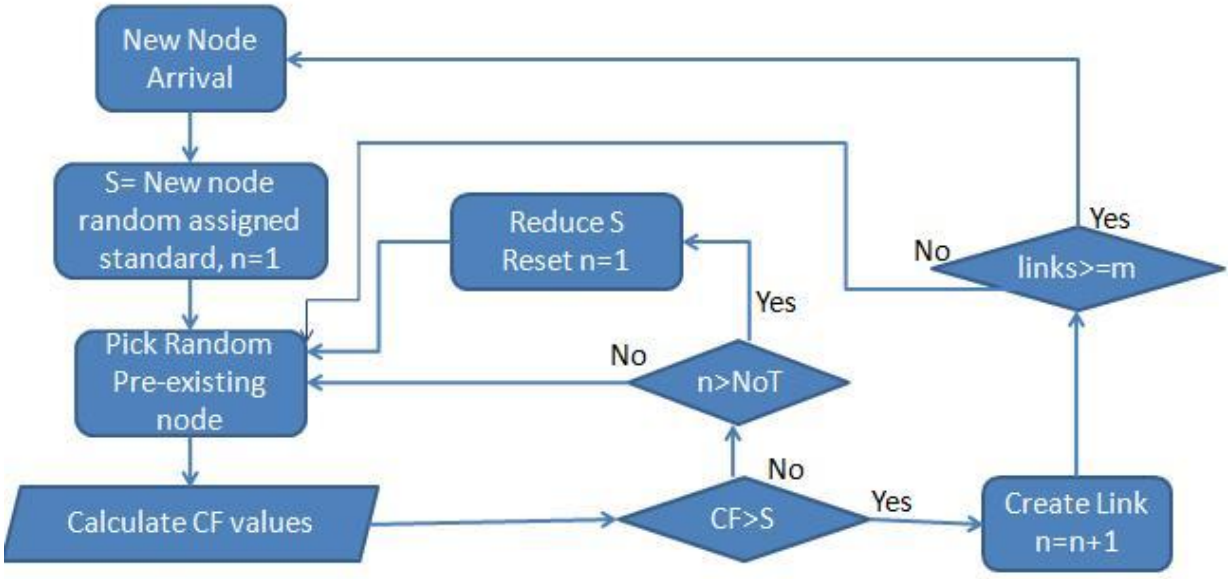
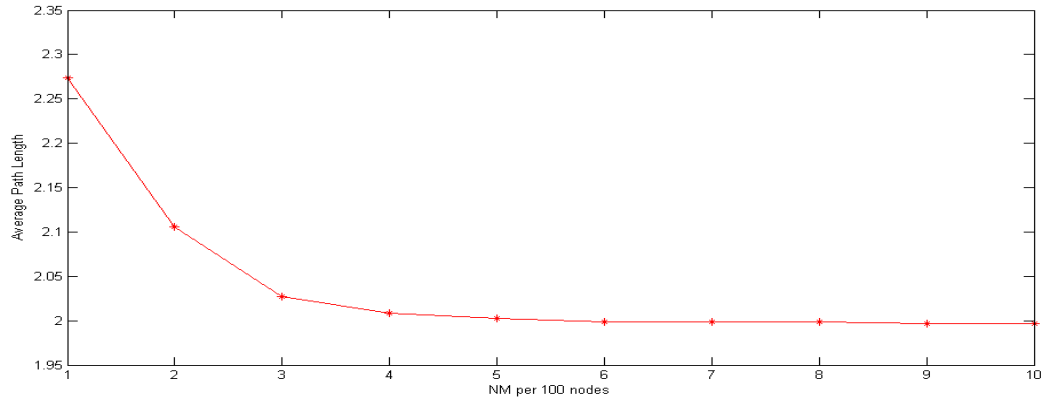


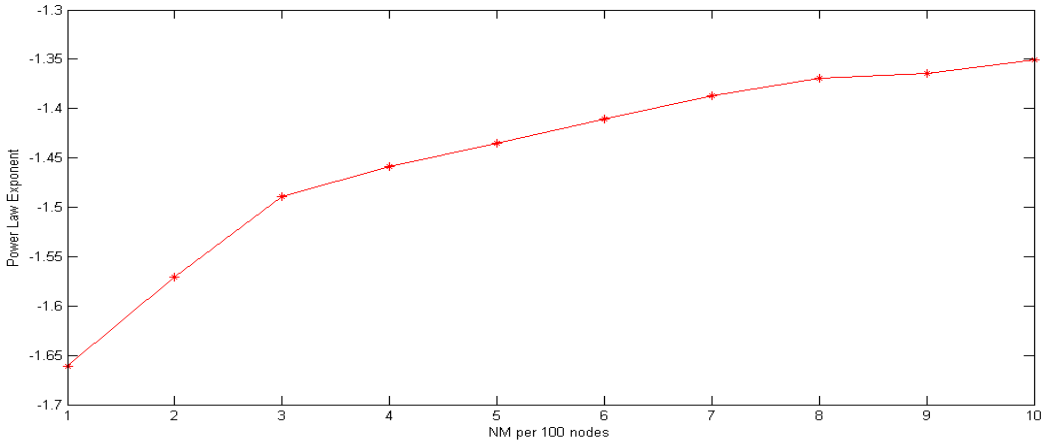
Figure 4.3. Flow chart of SNAM algorithm.

4.3.3. Simulation Results

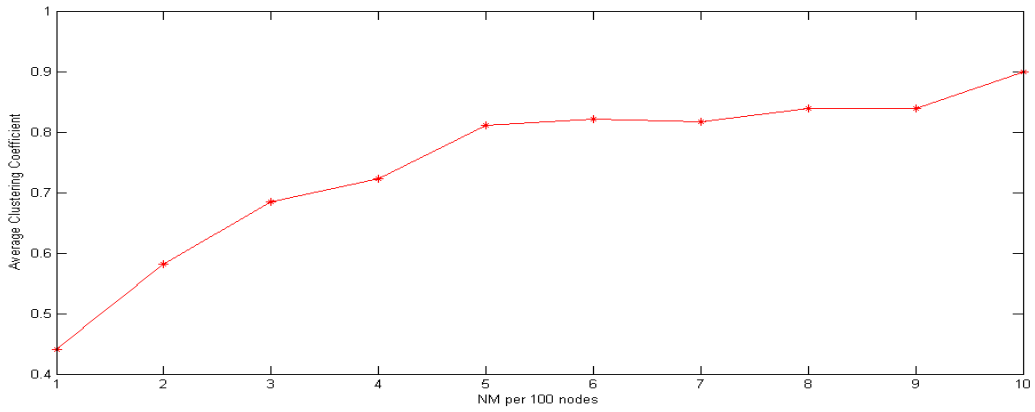
Results for three combinations of the coefficients α, β, and w are shown in Figures 4.4 through 4.6.



a) Average Path Length

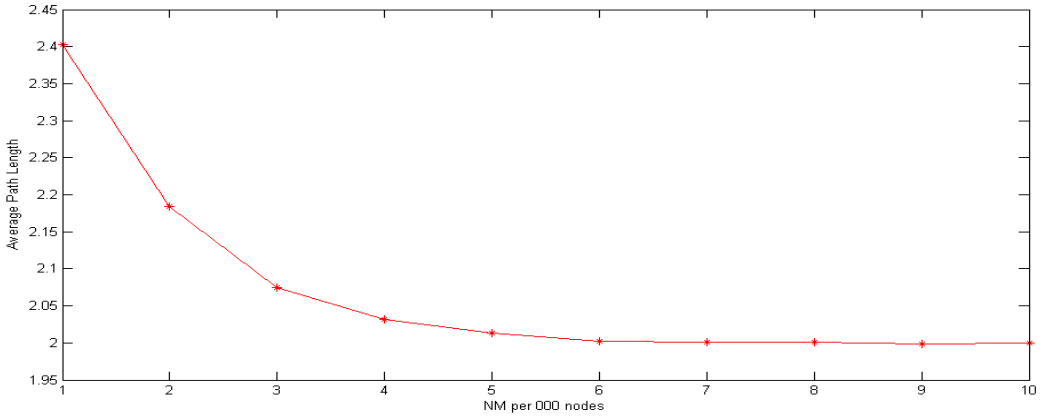


b) Power Law Exponent of Degree Distribution

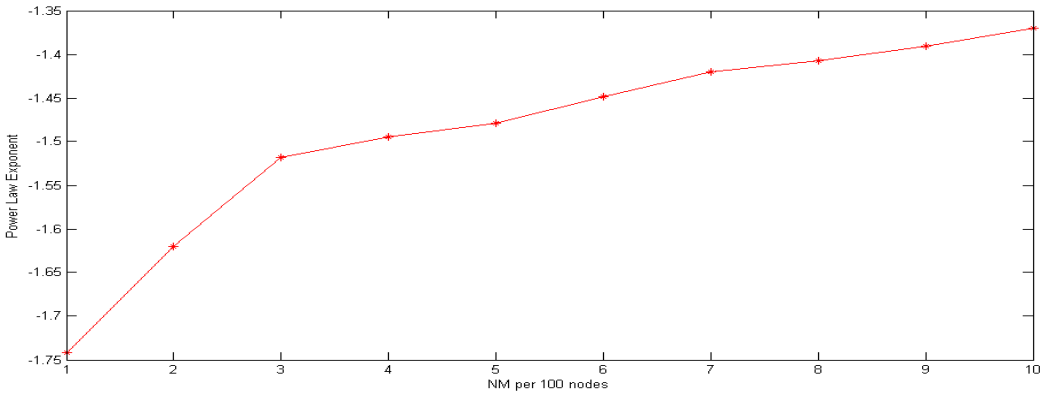


c) Average Clustering Coefficients

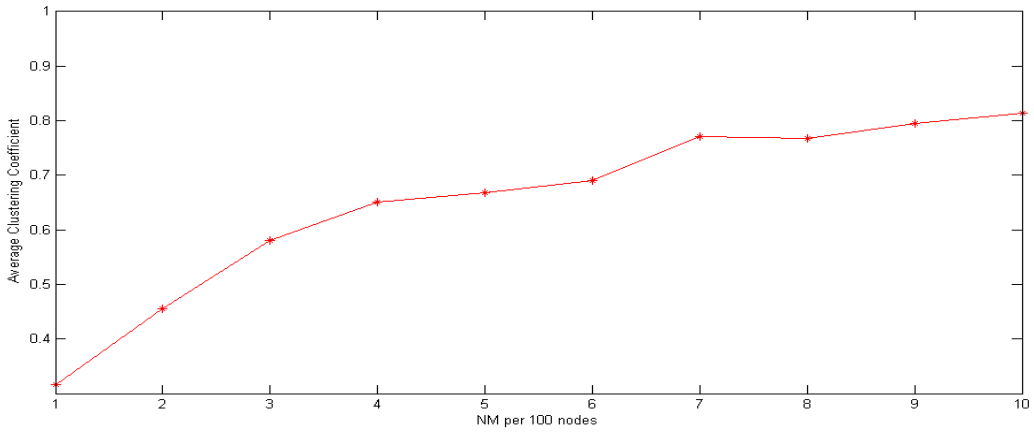
Figure 4.4. SNAM algorithm with a normalized degree *CF* (β = 1). a) Average Path Length, b) Power Law Exponent of Degree Distribution, c) Average Clustering Coefficients



a) Average Path Length

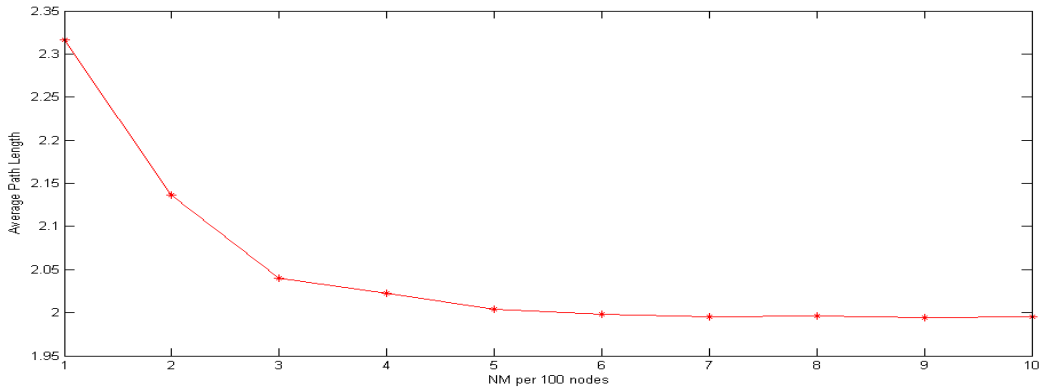


b) Power Law Exponent of Degree Distribution

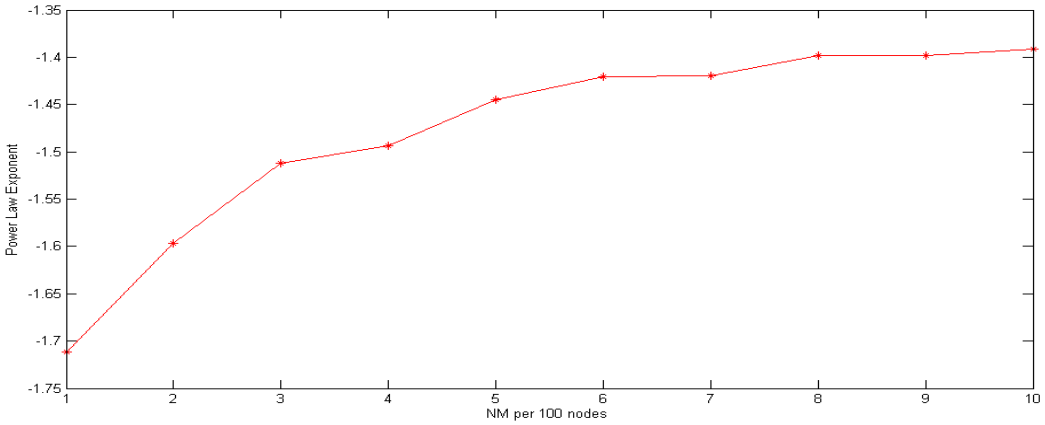


c) Average Clustering Coefficients

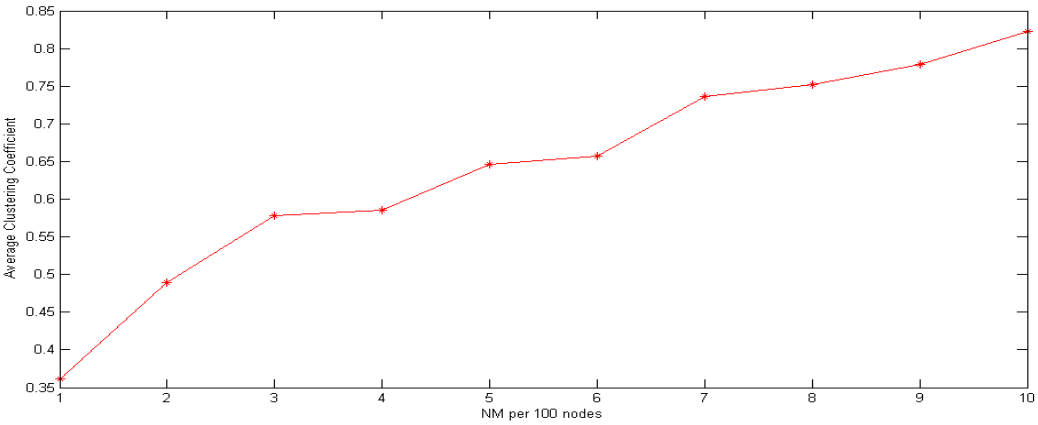
Figure 4.5. SNAM algorithm with a normalized degree with added attribute similarity *CF* ($w = \beta = 0.5$). a) Average Path Length, b) Power Law Exponent of Degree Distribution, c) Average Clustering Coefficients



a) Average Path Length



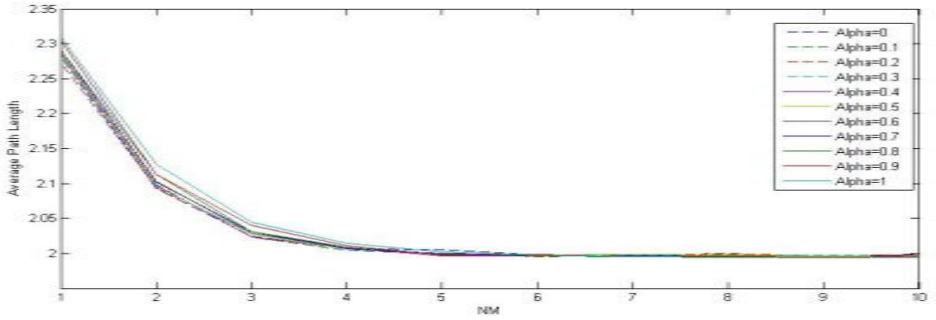
b) Power Law Exponent of Degree Distribution



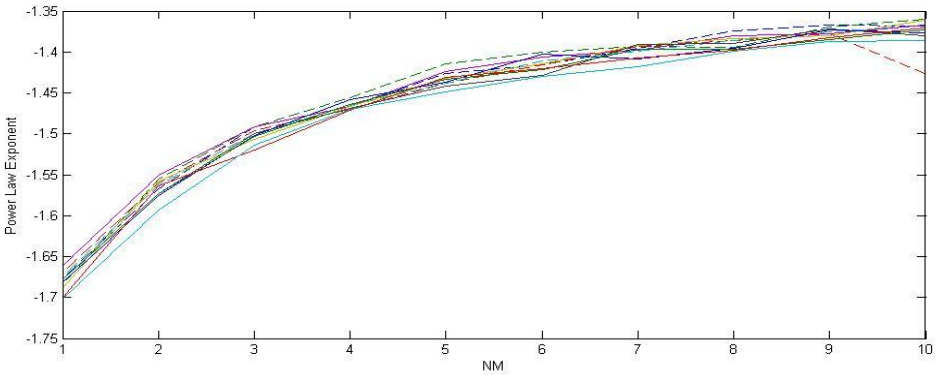
c) Average Clustering Coefficients

Figure 4.6. SNAM algorithm with a normalized degree with multiplied attribute similarity *CF* (α = 1). a) Average Path Length, b) Power Law Exponent of Degree Distribution, c) Average Clustering Coefficients

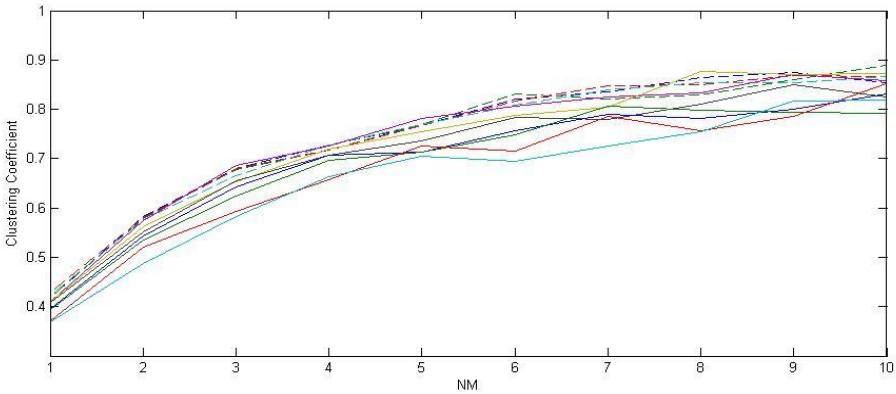
We also examine the effect that changing the values of the CF coefficients α and β has on the statistical properties of the generated network. Figure4.7 shows the effect of varying the coefficient of the multiplicative attribute similarity term α and that of the normalized degree β such that $\alpha = 1 - \beta$ and $w = 0$ on the resulting SNAM statistical properties.



a) Average Path Length



b) Power Law Exponent of Degree Distribution



c) Average Clustering Coefficients

Figure 4.7. SNAM algorithm with varying coefficient values for the normalized degree with multiplied attribute similarity *CF*. a) Average Path Length, b) Power Law Exponent of Degree Distribution, c) Average Clustering Coefficients

4.3.4. Analysis of Results

Figures 4.4 (a), 4.5 (a), and 4.6 (a) show the average path length. The results indicate that the small world effect is preserved for the three combinations of α, β and w. The average path length decreases with increasing *NM*. The average path length saturates at the value of 2 when $NM > 4$ for the three combinations of α, β and w. The resultant average path length for a network of size N generated by SNAM is less than that for the same sized network generated by IASM. The lower average path length of SNAM makes the networks generated using it more reliable and fault tolerant.

Figures 4.4 (b), 4.5 (b), and 4.6 (b) show that the magnitude of the PL exponents, γ, for the three variations remains in the range of $1.35 \leq \gamma \leq 1.75$, which is consistent with values found in real-world networks [2, 4, 5]. At *NM*=1, *NoT'*s value remains constant, the PL exponent values are 1.68, 1.72, and 1.69 for the graphs in Figures 4.4 (b), 4.5 (b), and 4.6 (b), respectively. Additionally, the magnitudes of the PL exponent saturate at values close to $\gamma \cong 1.35$ with increasing *NM*. This leads us to believe that the variation of the *CF* terms here had a minor effect on the obtained PL exponent values. However, increasing the *NM* values does lead to the decrease of γ magnitudes.

The average clustering coefficient values increase with increasing *NM* for the three variations of *CF* as shown in Figures 4.4 (c), 4.5 (c), and 4.6 (c). The average clustering coefficient reaches much higher values than those of the BA model or our IASMs. The clustering coefficients in Figure 4.4 (c), when *CF* depends on the test nodes' normalized degree value, achieve higher values than those shown in Figures 4.5 (c) and 4.6 (c) when the *CF* integrates the attribute similarity measure with the normalized degree. The choice of the *CF* coefficient values has a direct effect on the obtained average clustering coefficients unlike when we used TFS for increasing the average clustering coefficients in IASM.

Generally, the increase of *NM* led to the increase of higher degree nodes (hubs) which is reflected by the decrease of the magnitudes of γ. Additionally, the formation of hubs increases the triples in the network which can affect the obtained values of the average clustering coefficient. Results indicate that the number of average clustering coefficient values increases with the increase of *NM*. Also, having hubs in the network can be beneficial in decreasing the average path length as upon reaching a hub a network can reach many other nodes easily.

Figures 4.7 (a), 4.7 (b), and 4.7 (c) show the effect of varying the coefficient of the multiplicative attribute similarity term α and that of the normalized degree β in *CF* on resulting statistical properties of graphs generated using SNAM. When $\beta >> \alpha$, the effect of the *CF* term depending on β dominates and vice versa. Figure 4.7(a) shows that the higher the value of α (multiplicative attribute similarity), the higher the value of average path length. However, the small world phenomenon is preserved for all α and β values. The PL exponent values remain between $1.35 \leq \gamma \leq 1.75$ for all α and β values, as shown in Figure 4.7 (b). Figure 4.7 (c) also shows that the average clustering coefficient increases slightly with decreasing of α. Thus, the integration of attribute similarity into *CF* in SNAM has decreased the generated network's average clustering coefficient. However, this decrease is an acceptable trade off in order to include the idea that node attributes similarity can affect nodes' connections.

The SNAM generation model has preserved the PL degree distribution, has a small average path length, and has high clustering coefficient values. The value of parameter *NM* can be used to generate a variety of complex networks with specific values of the clustering coefficient, the average path length, and the PL exponent. For example even with the *CF* depending solely on the normalized degree, we can generate networks with different statistical properties. For *NM* = 2, the average path length equals 2.1, the PL exponent has the magnitude of 1.56, and the average clustering coefficient has the value of 0.68. Increasing *NM* to 4 gives the values of 2, 1.4, and 0.82 for the average path length, PL exponent, and the average clustering coefficient, respectively. Thus, we can tune the *NM* value that is used to generate a complex network with required specific statistical properties.

4.4. Community Structure in IASM and SNAM

4.4.1. Model Assumptions

One of the characteristics of a real-world network node is its belonging to a certain class. A node belonging to a certain class usually prefers to connect to nodes of a similar class. This preference leads to the existence of communities in the network where each community contains nodes of the same specific class. For example, a user's age in a social network can affect his or her social circle as most of their connections are made within their age group. Another example is a user's gender where females or males prefer to connect to users of the same gender.

We argue that including this preference in our network generation model can lead to the existence of community structure as found in real social networks. The community structure characteristic is present when actors have most of their connections with actors of their same class while only few connections link actors of different classes together. We deploy this in our model by including a parameter CS_{ij} that describes the class similarity between the added new node *i* and the old or existing test node *j* in the connection function, *CF*. This class similarity parameter takes the value of 1 if the arriving node and the test node belong to the same class and is 0otherwise. This gives nodes belonging to the same class a higher probability to get connected. Nodes from different classes can still be connected, but with a lower probability.

4.4.2. Simulation Setup and Parameters

These community structure models are modifications of the IASM and SNAM models. The connection function, *CF*, of the modified models includes the three terms for normalized node degree, for normalized attribute similarity, and for their normalized product in addition to an extra added class similarity term. One, two, or more nodes' attributes are used to compute class similarity between the nodes. This means having two, four, eight, or more classes. For arriving node *i* to form a connection with node *j*, the connection function, *CF*, is computed as:

$$\begin{aligned} \mathrm{CF} = (&\text{arr Normalized [(degree of node j)} \times \text{(Attribute Similarity between nodes i, j)]} \\ &+ (\beta) \times \text{ Normalized (degree of node j)} \\ &+ (w) \times \text{Normalized (attribute Similarity between nodes i, j)} \\ &+ (\mu) \times \text{(Class Similarity between nodes i, j)]}, \end{aligned} \tag{4.4}$$

where $\alpha + w + \beta + \mu = 1.0$, $0 \leq \alpha \leq 1$, $0 \leq w \leq 1$, and $0 \leq \beta \leq 1$.

4.4.3. Simulation Results

The following simulation results use one or two class similarity attributes corresponding to two or four classes. This avoids the complexity of class similarity computation. Tables 4.5 and 4.6 show the values for the percentage of inter-class connections among network connections (PERCENTAGE), magnitude for power law exponent (PL EXP), average clustering coefficient (AvClustr), and average path length (Av.PL) values of the generated networks using either IASM or SNAM. As indication of the community formation, Table 4.5 shows the percentage of

all inter-classes connections for networks generated for a subset of *CF* coefficients values α, w, β, and μ where $\alpha + w + \beta + \mu = 1$.

4.4.4. Analysis of Results

Increasing the weight of the class similarity coefficient μ in the CF has the following effects on the statistical properties of generated network.

- Increasing μ decreases the percentage of inter-class connections (connections between users belonging to different classes). Thus, enforcing the connections made between members of the same class (community). This emphasizes the presence of community structure within the network.
- The results show a small variation in the values of the magnitude of the generated PL exponent, but the values are still within the exponent values range reported in [2, 4, 5].
- Large values of class similarity coefficient μ increase the average path length to values slightly larger than 3. This is still smaller than 6, so the small world property is not lost. This increase in average path length is due to the addition of more constraints on the connections made between different users.
- An increase in the magnitude of PL exponent γ implies a decrease in number of higher degree nodes or hubs. This decrease in the number of hubs can cause an increase in the path lengths between nodes trying to reach each other since the hub could act as a relay for shorter connection between multiple nodes. Thus, the increase in average path length can be a result of the decrease in the number of hubs formed in the network.
- Increasing μ decreases the average clustering coefficient of the generated network in SNAM, but increases it slightly in IASM. We can see here further proof that the community structure and the average clustering coefficient are two different statistical properties unlike what some authors assume. Here, we can see that increasing the weight of the class similarity coefficient in SNAM strengthens the community structure and decreases the average clustering coefficient of the whole generated networks. This decrease in average clustering coefficient can also be tied to the decrease in the number of network hubs. This decrease in number of hubs has an effect on the number of triplets

formed in the network which is reflected in the value of the resultant average clustering coefficient of the network.

Table 4.5. Statistical Properties for 2-Class Networks Generated using SNAM and IASM

CF coefficient values				**SNAM**				**IASM**			
α	W	β	μ	PERCENTAGE	PL EXP.	Av.PL	AvClustr	PERCENTAGE	PL EXP.	Av.PL	AvClustr
1.0	0	0	0	23.6625	-1.4116	1.9949	0.8533	24.7080	-1.6726	3.2568	0.0145
.975	0	0	.025	17.2265	-1.3586	2.0516	0.6806	10.7303	1.6701	3.2746	0.0162
.95	0	0	.05	1.850	-4.6707	2.5918	0.4014	7.7001	-1.6823	3.2932	0.0180
.925	0	0	.075	0.3520	-3.3342	3.0090	0.2154	5.3856	-1.6737	3.3276	0.0198
.9	0	0	.1	0.1307	-2.6254	3.2993	0.1242	4.6132	-1.6846	3.3428	0.0215
.875	0	0	.125	0.1810	-2.2627	3.4334	0.0790	3.6829	-1.6596	3.3742	0.0214
.85	0	0	.15	0.2313	-2.1282	3.5570	0.0559	3.2772	-1.6496	3.3949	0.0220
.825	0	0	.175	0.1609	-2.0395	3.6554	0.0460	2.7886	-1.6464	3.4203	0.0228
.8	0	0	.2	0.2514	-1.932	3.681	0.0404	2.5227	-1.6711	3.4338	0.0228
0	0	1	0	24.8492	-2.0481	1.9989	0.7020	25.0686	-1.8359	3.1987	0.0170
0	0	.99	.01	23.1496	-1.5381	2.0174	0.6335	18.4877	-1.7781	3.2236	0.0162
0	0	.98	.02	22.7273	-1.8417	2.0411	0.6356	14.6094	-1.7496	3.2339	0.0162
0	0	.97	.03	22.5764	-1.3814	2.0578	0.6200	12.3513	-1.7330	3.2472	0.0165
0	0	.96	.04	19.4590	-1.4227	2.0761	0.6096	10.6119	-1.7248	3.2556	0.0170
0	0	.95	.05	17.3572	-1.3637	2.1069	0.5576	9.3388	-1.7311	3.2652	0.0174
0	0	.94	.06	13.3850	-3.0631	2.1505	0.5171	8.4054	-1.7099	3.2747	0.0185
0	0	.93	.07	6.2249	-4.3644	2.3534	0.4522	7.6876	-1.6851	3.2813	0.0185
0	0	.92	.08	3.6203	-4.5074	2.6471	0.3946	6.9538	-1.6952	3.2879	0.0191
0	0	.91	.09	0.2615	-4.1780	2.8461	0.3364	6.3849	-1.7123	3.2970	0.0203
0	0	.9	.1	0.3017	-3.7224	2.9467	0.2795	5.9105	-1.7014	3.3082	0.0196
0	.5	.5	0	25.2313	-1.4768	2.0001	0.6463	22.7602	-1.7573	3.2458	0.0151
0	.49	.49	.02	20.4143	-1.3225	2.0416	0.6563	21.3183	-1.7236	3.2499	0.0148
0	.48	.48	.04	2.0414	-4.5022	2.4789	0.4477	20.0122	-1.7131	3.2519	0.0147
0	.47	.47	.06	0.3017	-3.5640	2.9787	0.2405	18.6434	-1.7243	3.2496	0.0152
0	.46	.46	.08	0.3117	-2.8104	3.1898	0.1393	17.7206	-1.7263	3.2538	0.0149
0	.45	.45	.1	0.1911	-2.5131	3.3855	0.0878	16.5718	-1.7098	3.2547	0.0184
0	.44	.44	.12	0.2212	-2.2862	3.4751	0.0663	15.4229	-1.6991	3.2585	0.0147
0	.43	.43	.14	0.2212	-2.1249	3.5456	0.0499	14.6329	-1.6979	3.2604	0.0151
0	.42	.42	.16	0.2212	-2.0136	3.6184	0.0431	13.6072	-1.7316	3.2594	0.0155
0	.41	.41	.18	0.1911	-1.9702	3.6460	0.0380	13.1421	-1.6905	3.2637	0.0152
0	.4	.4	.2	0.1710	-1.9045	3.6950	0.0348	12.5000	-1.6779	3.2676	0.0152

Table 4.6. Statistical Properties for 4-Class Networks Generated using SNAM and IASM

CF coefficient values				**SNAM**				**IASM**			
A	W	β	μ	PERCENTAGE	PL EXP.	Av.PL	Av.Clustr	PERCENTAGE	PL EXP.	Av.PL	Av.Clustr
1.0	0	0	0	36.2531	-1.3294	1.9936	0.7627	37.0709	-1.6691	3.2556	0.0142
.975	0	0	.025	24.9397	-1.3453	2.0571	0.6598	22.4373	-1.6671	3.2726	0.0162
.95	0	0	.05	2.3933	-2.9327	2.9568	0.2172	17.6844	-1.7010	3.2853	0.0189
.925	0	0	.075	.5631	-2.3084	3.5813	0.1176	13.6476	-1.6529	3.3180	0.0233
.9	0	0	.1	.6537	-2.0715	3.7893	0.0865	11.9294	-1.6909	3.3379	0.0251
.875	0	0	.125	.5733	-1.9607	3.8958	0.0710	9.8489	-1.6790	3.3740	0.0288
.85	0	0	.15	.6034	-1.9008	3.9266	0.0664	8.7999	-1.6731	3.3963	0.0300
.825	0	0	.175	.6536	-1.8766	3.9735	0.0634	7.4743	-1.6682	3.4355	0.0328
.8	0	0	.2	.4325	-1.8207	4.0214	0.0602	6.9119	-1.6649	3.4506	0.0334
0	0	1	0	37.4496	-2.3210	1.9975	0.7061	37.5077	-1.8169	3.2002	0.0171
0	0	.99	.01	35.3781	-2.7901	2.0167	0.7054	32.006	-1.8059	3.2107	0.0165
0	0	.98	.02	33.7287	-2.1602	2.0288	0.7295	27.9235	-1.7345	3.2277	0.0161
0	0	.97	.03	31.5664	-2.0135	2.0450	0.6182	24.8246	-1.7604	3.2316	0.0172
0	0	.96	.04	29.8572	-2.0520	2.0707	6170	22.1974	-1.7393	3.2400	0.0176
0	0	.95	.05	26.7700	-3.5844	2.0902	0.5454	20.2097	-1.7548	3.2499	0.0185
0	0	.94	.06	22.9184	-4.0631	2.1784	0.4489	18.6136	-1.7328	3.2571	0.0198
0	0	.93	.07	14.7526	-3.4720	2.3151	0.3577	17,1202	-1.7429	3.2682	0.0204
0	0	.92	.08	7.6126	-2.7553	2.8337	0.2006	15.9947	-1.7234	3.2773	0.0211
0	0	.91	.09	2.2124	-2.5554	3.2124	0.1408	12.3491	-1.7125	3.2856	0.0215
0	0	.9	.1	1.0760	-2	3.4094	0.1232	14.0462	-1.7036	3.2973	0.0228
0	.5	.5	0	37.4497	-1.6435	2.0004	0.6328	35.3259	-1.7244	3.2467	0.0144
0	.49	.49	.02	30.6417	-1.5473	2.0342	0.6380	34.0542	-1.7492	3.2470	0.0147
0	.48	.48	.04	5.4204	-2.9574	2.7152	0.2354	33.0114	-1.7262	3.2496	0.0148
0	.47	.47	.06	0.4928	-2.3661	3.5973	0.1131	31.8473	-1.7276	3.2485	0.0150
0	.46	.46	.08	0.6236	-2.1388	3.7664	0.0850	30.7919	-1.7284	3.2481	0.0149
0	.45	.45	.1	0.6235	-2.0056	3.8470	0.0731	29.7874	-1.7092	3.2537	0.0146
0	.44	.44	.12	0.4627	-1.9415	3.9120	0.0672	28.6919	-1.7148	3.2517	0.0148
0	.43	.43	.14	0.5129	-1.8647	3.9488	0.0621	27.7111	-1.7110	3.2550	0.0153
0	.42	.42	.16	0.5229	-1.8223	3.9784	0.0616	23.8104	-1.7188	3.2569	0.0150
0	.41	.41	.18	0.6035	-1.8318	4.0196	0.0601	25.7385	-1.7266	3.2554	0.0156
0	.4	.4	.2	0.5028	-1.7855	4.0414	0.0579	24.6436	-1.7056	3.2598	0.0161

- Additionally, including the attribute similarity whether multiplicative (α) or additive (w) led to a faster decrease in the percentage of inter-class connections with increasing μ. This can be the result of our choice of dividing the network into different classes according to some (1, 2, or more) of the 10 nodes' attributes used in the simulation.
- It is clear from different results that to reach the same level of community structure formation, IASM requires higher values of class similarity coefficient than those required by SNAM.

Our IASM and SNAM community structure models preserved PL degree distribution that characterize real-world networks. The small world property is also preserved. The models show community structure as most of the connections are made between nodes belonging to the same class with only a small percentage of the connections made between nodes of different classes.

Chapter 5. Mathematical Analysis of IASM and SNAM

5.1. Introduction

In this chapter, we analytically study IASM and SNAM. The rate equation analysis method and mean field theory are used to show that our IASM and SNAM models have a node degree distribution that follows a power law distribution. The study presents expressions for the power law exponent under certain model conditions.

5.2. IASM Analysis

5.2.1. Introduction

The Integrated Attribute Similarity Model, IASM_A, follows the same connection algorithm as the Barabási-Albert model [3]. As in the BA model, IASM models a growing network where connections are made using preferential attachment. Simulation results, presented in Chapter 4, show that IASM preserves the power law degree distribution found in real-world complex networks. IASM's connection function, CF, depends on both the existing nodes' degrees and the attribute similarity between the newly added node and existing network nodes. To verify the power law dependence found through simulation and to better understand IASM, the analysis presented in this chapter derives an expression for the node degree distribution using the rate equation analysis method and mean field theory [42, 1]. We concluded from our simulation results that, for IASM, the multiplicative attribute similarity measure is a better measure for similarity than the additive attribute similarity. Therefore, the analysis here is done for IASM with multiplicative attribute similarity.

Both the BA model and IASM model start network growth with a seed network with m_o nodes connected by e_o edges given that $m_o \geq 2$ and $e_o \geq 1$. At each time step, a new node l is added to the network with m links, $m \leq m_o$, to be connected one by one preferentially to the existing network old nodes. The addition of new nodes continues until the final network size is reached.

5.2.2. BA Model Degree Distribution

We begin by repeating the derivation for the degree distribution performed by Barabási and Albert [3]. We then use this derivation as a guide for our mathematical analysis of IASM. The connection function used in the BA model, *CF*, which is used in the BA preferential attachment

(PA) connection algorithm, depends only on an existing node's degree k_i. Its rate equation is given by:

$$\frac{\partial k_i(t)}{\partial t} = m\frac{k_i}{\sum_j k_j} = m\frac{k_i(t)}{2e_o+2mt}$$

Since, for large values of *t*, $2mt \geq e_o$[42],

$$\frac{\partial k_i(t)}{\partial t} \cong \frac{1}{2}\frac{k_i(t)}{t} \tag{5.1}$$

Integration gives, $\ln\frac{k_i(t)}{k_i(t_i)} = \frac{1}{2}\ln\frac{t}{t_i}$.

Since node *i* was created or born at t_i, $k_i(t_i) = m$. Therefore,

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{1/2} \tag{5.2}$$

Using $P(k) = \frac{\partial P[k_i(t)<k]}{\partial k}$ [42] and Equation 5.2, Barabási and Albert prove that the node degree distribution follows a power law, $P(k) \propto k^{-\gamma}$, having exponent $\gamma = 3$ [42].

5.2.3. IASM Model with Multiplicative Attribute Similarity Analysis

As previously mentioned, we consider the IASM model case where *CF* depends on the normalized existing node degree multiplied by the value of the normalized attribute similarity. The normalized attribute similarity between existing node *i* and the new arriving node *j*, S_{ij}, is computed as:

$$S_{ij} = \frac{1}{L}\left\|\begin{bmatrix}a_{i1}\\a_{i2}\\\vdots\\a_{iL}\end{bmatrix} \circ \begin{bmatrix}a_{j1}\\a_{j2}\\\vdots\\a_{jL}\end{bmatrix}\right\| = \frac{1}{L}\sum_{l=1}^{L} a_{il}a_{jl} = \frac{1}{L}\sum_{l=1}^{L} b_{ijl}$$

Operator (°) represents the element-by-element multiplication of the attribute vector $\begin{bmatrix}a_{i1}\\a_{i2}\\\vdots\\a_{iL}\end{bmatrix}$ of node *i* and the attribute vector $\begin{bmatrix}a_{j1}\\a_{j2}\\\vdots\\a_{jL}\end{bmatrix}$ of node *j*. Division by *L* normalizes S_{ij}.

Next, we use the probability density functions $\rho(a_{il})$ and $\rho(a_{jl})$ to determine the probability density functions, $\rho(b_{ijl})$ and $\rho(S_{ij})$. The random variables a_{il}*'s* are assumed to be independent for different values of l and each random variable a_{il} is binary, either 1 or 0, both with equal probability, as shown in Figure5.1 (a).

(a) $\rho(a_{il})$ (b) $\rho(b_{ijl})$

Figure 5.1. Probability density functions, $\rho(a_{il})$ and $\rho(b_{ijl})$.

Then,

$P(a_{il} = 1) = \frac{1}{2} = P(a_{il} = 0) = \frac{1}{2}$ or $\rho(a_{il}) = \frac{1}{2}\delta(a_{il}) + \frac{1}{2}\delta(a_{il}\text{-}1)$,

$\rho(a_{jl}) = \frac{1}{2}\delta(a_{jl}) + \frac{1}{2}\delta(a_{jl}\text{-}1)$

Thus, as shown in Figure 5.1 (b), $\rho(b_{ijl}) = \rho(a_{il}a_{jl}) = \frac{3}{4}\delta(b_{ijl}) + \frac{1}{4}\delta(b_{ijl}\text{-}1)$.

Since $S_{ij} = \frac{1}{L}\sum_{l=1}^{L} b_{ijl}$ and all b_{ijl}'s are independent for all values of l, the random variable $S_{ij} = \frac{1}{L}\{0\ or\ 1\ or\ 2\ \ldots\ldots or\ L\}$ and its pdf follow a Binomial distribution.

Using the notation $p = P(b_{ijl} = 1) = \frac{1}{4}$, then $(1 - p) = P(b_{ijl} = 0) = \frac{3}{4}$, and the pdf of S_{ij} is:

$P(S_{ij}=\frac{w}{L}) = \frac{L!}{(L-w)!\ w!}p^{w}(1 - p)^{L-w}$, where w is a constant and $w = L, L\text{-}1,\ldots, 3, 2, 1, 0$.

Substituting $p = \frac{1}{4}$ gives

$P(S_{ij} = \frac{w}{L}) = \frac{L!}{(L-w)!\ w!}(\frac{1}{4})^{w}(\frac{3}{4})^{L-w} = \frac{L!}{(L-w)!\ w!}(\frac{1}{4})^{L}(3)^{L-w}$

Hence, the probability density function, $\rho(S_{ij})$ is:

$$\rho(S_{ij}) = \sum_{w=0}^{L} \frac{L!}{(L-w)!\ w!}(\frac{1}{4})^{L}(3)^{L-w}\,\delta(S_{ij}\text{-}\frac{w}{L}) \qquad (5.3)$$

Since the IASM model, as considered here, has *CF* dependent on the normalized node degree of existing node *I* multiplied by the value of the normalized attribute similarity between the existing node *i* and the new arriving node *j*, its rate equation is given by:

$$\frac{\partial \mathrm{k}_i(\mathrm{t})}{\partial \mathrm{t}} = \mathrm{m}\frac{\mathrm{k}_i S_{ij}}{\sum_j \mathrm{k_j} S_{ij}}$$

Using mean field theory [1, 42], $\sum_{\mathrm{j}} \mathrm{k_j}\, \mathrm{S_{ij}}$ is replaced by its mean value, $\langle \sum_j \mathrm{k}_j S_{ij} \rangle$.

Therefore,

$$\frac{\partial \mathrm{k_i(t)}}{\partial \mathrm{t}} \cong \mathrm{m}\frac{\mathrm{k}_i S_{ij}}{\langle \sum_j \mathrm{k_j} S_{ij} \rangle} = \mathrm{mt}\frac{S_{ij}}{\langle \sum_j \mathrm{k_j}\, S_{ij} \rangle}\frac{\mathrm{k}_i}{t}$$

$$\frac{\partial \mathrm{k_i(t)}}{\partial \mathrm{t}} = \beta(\mathrm{S_{ij}})\frac{\mathrm{k}_i}{t} \tag{5.4}$$

Similar to the scale-free BA model, comparing Equations 5.1 and 5.4 indicates that it is assumed that the degree of node *i* at time evolution $\mathrm{k}_i(t)$ follows a power law:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta(S_{ij})} \tag{5.5}$$

From Equation 5.4,

$$\beta(S_{ij}) = mt\frac{S_{ij}}{\langle \sum_j k_{\mathrm{j}}\, S_{ij} \rangle} \tag{5.6}$$

Here, $m = k_i(t_i)$ = degree of node *i* at birth

Equation 5.5 indicates that the growth of the node degree in IASM is a multi-scaled system with a dynamic exponent dependent on S_{ij}. Notice that the dynamic exponent $\beta\left(S_{ij}\right)$ in the expression for $k_{\mathrm{i}}(t)$, Equation 5.5, must be positive, $\beta\left(S_{ij}\right) > 0$, so the node's degree at different time steps either remains constant or increases. Also,since a node's degree cannot grow faster than evolution time *t*, the exponent cannot exceed 1, $\beta\left(S_{ij}\right) < 1$. Therefore, both conditions imply that

$0 < \beta\left(S_{ij}\right) < 1.$

To study this dynamic exponent, $\beta\left(S_{ij}\right)$, its denominator is manipulated as follows:

$$\langle \sum_{\mathrm{i}} \mathrm{k_i S_{ij}} \rangle = \int_{\mathrm{S_{ij}}_{\min}}^{\mathrm{S_{ij}}_{\max}} \mathrm{S_{ij}}\rho\left(\mathrm{S_{ij}}\right) \int_{\mathrm{t_i}=1}^{\mathrm{t}} \mathrm{k_i(t)}\, \mathrm{dt_i}\, \mathrm{dS_{ij}},\ \mathrm{S_{ij}} \neq 0$$

$$= \int_{S_{ij_{min}}}^{1} S_{ij}\,\rho(S_{ij}) \int_{t_i=1}^{t} m\left(\frac{t}{t_i}\right)^{\beta(S_{ij})} dt_i\ dS_{ij}$$

$$= m\,t \int_{S_{ij_{min}}}^{1} S_{ij}\rho(S_{ij})\,(t)^{\beta(S_{ij})-1} \int_{t_i=1}^{t} (t_i)^{-\beta(S_{ij})}\ dt_i\ dS_{ij}$$

$$= m\,t \int_{S_{ij_{min}}}^{1} S_{ij}\rho(S_{ij})\,(t)^{\beta(S_{ij})-1} \frac{[t^{1-\beta(S_{ij})}-1]}{1-\beta(S_{ij})} dS_{ij}$$

$$= m\,t \int_{S_{ij_{min}}}^{1} \frac{S_{ij}\rho(S_{ij})}{1-\beta(S_{ij})}[1 - t^{-(1-\beta(S_{ij}))}]dS_{ij} \quad (5.7)$$

Since $0 < \beta(S_{ij}) < 1$, $\left(1 - \beta(S_{ij})\right) > 0$. Therefore, the expression for $\langle \sum_i k_i S_{ij} \rangle$can be written as:

$$\langle \sum_i k_i S_{ij} \rangle = Cmt[1 + O(t^{-\varepsilon})] \quad (5.8)$$

where, $\varepsilon = [1 - \max(\beta(S_{ij})]$ is positive.

Therefore, as $t \to \infty$, Equations 5.7 and 5.6 give

$$\langle \sum_i k_i S_{ij} \rangle \cong C\,m\,t,\ \beta(S_{ij}) \cong \frac{S_{ij}}{C} \quad (5.9)$$

Thus, the degree time evolution power law, Equation 5.5 in terms of *C* is:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{S_{ij}}{C}} \quad (5.10)$$

Constant *C* is the solution of $C = \int_{S_{ij_{min}}}^{1} \frac{S_{ij}\rho(S_{ij})}{1-\frac{S_{ij}}{C}} dS_{ij}$,

where, $0 < \frac{S_{ij}}{C} < 1$ for all S_{ij}. and thus $w = L, L\text{-}1, \ldots, 3, 2, 1$.

Hence,

$$0 < (S_{ij})_{max} = 1 < C \quad (5.11)$$

Using the expression for $k_i(t)$, Equation 5.10, and the value of *C* from the solution of Equation 5.11, we can derive the probability distribution function for the degree $P[k]$. We replace the random variable $S_{ij}$by the dummy variable *S* for simplicity. The degree probability density function for a specific value of attribute similarity *S* is given by

$\rho_s(k) = \frac{\partial P[k_i(t)<k]}{\partial k}$.

From Equation 5.10, $P[k_i(t) < k] = P[m\left(\frac{t}{t_i}\right)^{\frac{S}{C}} < k] = P[t_i > t m^{\frac{C}{S}} k^{-\frac{C}{S}}] = 1 - P[\, t_i \leq t\, m^{\frac{C}{S}} k^{-\frac{C}{S}}]$

Since one vertex is created at each time instant,

$P[k_i(t) < k] = [1-\frac{t\, m^{C/S} k^{-C/S}}{m_o + t}]$

Then, $P[k_i(t) < k] \cong [1- m^{C/S} k^{-C/S}]$ for large values of *t* and for $S \neq 0$.

Giving, $\rho_s(k) = \frac{\partial P[k_i(t)<k]}{\partial k} = \frac{C}{S}\, m^{C/S} k^{-(\frac{C}{S}+1)}$, $S \neq 0$.

This is for a specific value of attribute similarity *S*, hence,

$$\rho(k) = \int_{S_{min}}^{1} \rho_s(k)\, \rho(S) dS = \int_{S_{min}}^{1} \frac{C}{S} m^{C/S}\, \rho(S) k^{-\left(\frac{C}{S}+1\right)} dS, S \neq 0. \quad (5.12)$$

Therefore, to find ρ(K), Equations 5.11 and 5.12 are used, rewritten as:

$$C = \int_{S_{min}}^{1} \frac{S\rho(S)}{1-\frac{S}{C}} dS \quad (5.11)^*$$

$$\rho(K) = \int_{S_{min}}^{1} \frac{C}{S}\, m^{C/S}\, \rho(S)\, k^{-\left(\frac{C}{S}+1\right)}\, dS, S \neq 0. \quad (5.12)^*$$

For IASM, $\rho(S)$ used in Equations 5.11 and 5.12, is shown previously in Equation 5.3 to follow a binomial distribution as:

$\rho(S) = \sum_{w=0}^{L} F(w)\delta(S_{ij} - \frac{w}{L})$, where $F(w) = \frac{L!}{(L-w)!w!}(\frac{1}{4})^L (3)^{L-w}$

for which, $C = \int_{S_{min}}^{1} \frac{S \sum_{w=0}^{L} F(w)\delta(S-\frac{w}{L})}{1-\frac{S}{C}} dS$, $S \neq 0$ and S_{min} corresponds to *w*=1.

Integration gives,

$$C = \sum_{w=1}^{L} \frac{\frac{w}{L}F(w)}{1 - \frac{w/L}{C}} = \sum_{w=1}^{L} \frac{\frac{w}{L}F(w)}{1 - \frac{w}{LC}},\ C > (S_{ij})_{max} = 1.0 \quad (5.13)$$

Equation 5.12* becomes

$\rho(K) = \int_{S_{min}}^{1} \frac{C}{S} m^{C/S} k^{-(\frac{C}{S}+1)} \sum_{w=0}^{L} F(w)\delta(S-\frac{w}{L})\, dS$, $S \neq 0$.

This integration has a value only at $S = \frac{w}{L}$:

$$\rho(K) = \sum_{w=1}^{L} \frac{C}{w/L} m^{\left(\frac{C}{w/L}\right)} \; k^{-\left(\left(\frac{C}{w/L}\right)+1\right)} \; F(w)$$

$$= \sum_{w=1}^{L} \left\{ \frac{LC}{w} \; m^{\left(\frac{LC}{w}\right)} F(w) \right\} k^{-\left(\left(\frac{LC}{w}\right)+1\right)} \qquad (5.14)$$

Thus, $P(k)$ is the sum of PL degree distributions having exponents

$\gamma_w = \left(\frac{LC}{w}\right) + 1$, where, $w = L, L\text{-}1, \ldots, 3, 2, 1$.

Thus, $\gamma_w = -(1+C), -\left(1+C\frac{L}{L-1}\right), -\left(1+C\frac{L}{L-2}\right), \ldots, -(1+LC)$ (5.15)

for each value of C given by Equation 5.13.

5.2.4. Numerically Finding Constant Values of *C*

In this section, C is determined through a numerical solution of Equation5.13 and the use of the condition, $C > (S_{ij})_{max} = 1.0$. Values of C for different values of L, the number of node attribute values, are found using the corresponding closed form expression for the degree's probability distribution function, $\rho[k]$.

5.2.4.1. Single Attribute Node, *L* = 1

If nodes have only one attribute, $L = 1$:

$S = 0$ or 1 and $\rho(S) = \frac{3}{4}\delta(S) + \frac{1}{4}\delta(S-1)$

Thus, $F(0) = \frac{3}{4}$ and $F(1) = \frac{1}{4}$ and Equation 5.13 gives

$C = \left(\frac{1}{4}\right)^{\frac{1}{1-\frac{1}{C}}}$, giving $C = 1.25 > [(S_{ij})_{max} = 1.0]$

And Equation 5.14 for $L = 1$ implies

$\rho[k] = \frac{1}{4} C \, m^{C} k^{-((C)+1)} = \frac{5}{16} \, m^{(1.25)} k^{-(2.25)}$

Thus, the degree distribution for $L = 1$ follows a power law with $\gamma = 2.25$.

5.2.4.2. Double Attribute Node, *L*=2

If nodes have two attributes, *L*=2:

For $L = 2$, $S = 0, \frac{1}{2}$ or 1, hence $w = 0$, 1, or 2 and $F(0) = \frac{9}{16}$, $F(1) = \frac{6}{16}$ and $F(2) = \frac{1}{16}$ and Equation 5.3 gives: $\rho(S) = \frac{9}{16}\delta(S) + \frac{6}{16}\delta(S-\frac{1}{2}) + \frac{1}{16}\delta(S - 1)$.

Thus, Equation 5.13 gives

$$C = \frac{\frac{1}{2}F(1)}{1-\frac{1}{2C}} + \frac{1\,F(2)}{1-\frac{1}{C}} = \frac{\frac{1}{2}\times\frac{6}{16}}{1-\frac{1}{2C}} + \frac{1\times\frac{1}{16}}{1-\frac{1}{C}}$$

Thus, $16 = \frac{6}{2C-1} + \frac{1}{C-1}$

Giving $32\ C^2 - 56\ C + 23=0$, whose roots are $C = 1.09 > [(S_{ij})_{max} = 1.0]$ which is accepted according to the condition of Equation 5.11, and $C = 0.6585 < [(S_{ij})_{max} = 1.0]$ which is not accepted according to the condition of Equation 5.11.

Also, for $L = 2$, Equation 5.14 becomes:

$$P[k] = \sum_{w=1}^{2} \frac{2C}{w} m^{\left(\frac{2C}{w}\right)} k^{-\left(\left(\frac{2C}{w}\right)+1\right)} [F(1)\ \delta(w-1) + F(2)\ \delta(w-2)]$$

$$= \frac{C}{1/2} m^{\left(\frac{C}{1/2}\right)} k^{-\left(\left(\frac{2C}{1}\right)+1\right)} F(1) + \frac{C}{2/2}\ m^{\left(\frac{C}{2/2}\right)} k^{-\left(\left(\frac{2C}{2}\right)+1\right)} F(2)$$

$$= 2C\ m^{(2C)} k^{-((2C)+1)} \times \frac{6}{16} + C\ m^{(C)} k^{-(C+1)} \times \frac{1}{16}$$

$$= \frac{3}{4} C\ m^{(2C)} k^{-((2C)+1)} + \frac{1}{16} C\ m^{(C)} k^{-(C+1)}$$

Therefore, using the accepted root $C = 1.0915 > [(S_{ij})_{max} = 1.0]$:

$$P[k] = \frac{3.2745}{4}\ m^{(2.183)} k^{-(3.183)} + \frac{1.0915}{16} m^{(1.0915)} k^{-(2.0915)},$$

Since for large values of *k*, $k^{-(2.09)} >> k^{-(3.18)}$;

Thus, for $L = 2$, $C = 1.0915$ is the only accepted value of C and the PL exponent $\gamma = 2.0915$ overpowers the one with $\gamma = 3.18$ and dominates the value of the summation.

5.2.4.3. Multiple Attribute Node

As seen for $L = 1$ and $L = 2$, it was possible to solve Equation 5.13 analytically to find values of C. However, finding an analytical solution for Equation 5.13 at higher values of L is challenging. Therefore, Equation 5.13 is numerically solved using MATLAB for values of L greater than 2. Table 5.1 shows the results found by numerical solution for $L = 3$, $L = 4$ and $L = 5$.

As mentioned before, there are L corresponding values of γ for each of the values of C, as shown in Table 5.2. Thus, there are L values of γ corresponding to the constant C satisfying the condition that $C > 1$. The smallest of these L values is the dominant exponent. It is clear from Table 5.2 that we can use IASM to generate complex networks with different power law exponent values. Thus, according to the desired power law exponent, we can vary the number of attributes assigned per node, L.

Table 5.1. Root Values C of Equation 5.13

L	C_1	C_2	C_3	C_4	C_5
1	1.25				
2	0.6585	1.0915			
3	0.4319	0.7891	1.029		
4	0.3138	0.6092	0.8194	1.0076	
5	0.243	0.4861	0.6902	0.829	1.0018

Table 5.2. Power Law Exponents for Different Values of L and Corresponding Values of C

L	C	γ_1	γ_2	γ_3	γ_4	γ_5
1	1.25	2.25				
2	0.6858	1.6585	2.317			
	1.0915	2.0915	3.183			
3	0.4319	1.4319	1.6419	2.2957		
	0.7891	1.7891	2.1837	3.3673		
	1.029	2.029	2.5435	4.0870		
4	0.3138	1.3138	1.418	1.6276	2.2552	
	0.6092	1.6092	1.8123	2.2184	3.4368	
	0.8194	1.8194	2.0925	2.6388	4.2776	

	1.0076	2.0076	2.3435	3.0152	5.0304	
5	0.243	1.243	1.3037	1.4050	1.6075	2.2150
	0.4861	1.4861	1.6076	1.8102	2.2152	3.4305
	0.6902	1.6902	1.8628	2.1503	2.7255	4.4510
	0.829	1.829	2.0360	2.3817	3.0725	5.1450
	1.0018	2.0018	2.2523	2.6697	3.5045	6.009

Next, we study which of the L values of γ corresponding to $C > 1$ is dominant and overpowers the summation $P(k)$. To decide what will happen if L has very large values, we go back to IASM with $L = 2$. It was previously stated that $P(k)$ for $\gamma = 2.0915$ (corresponding to $C > 1$ value of $C = 1.0915$) seems to dominate the summation of the two power law distributions. To test this, $P_1(k)$ and $P_2(k)$ that correspond to each of the two values of γ values (2.0915 and 3.183) as well as the sum of both, total $P(k)$ were plotted. As shown in Figure 5.2, it is clear that, for high values of k, $P_1(k)$ corresponding to lower value of γ is closer to total $P(k)$ and is dominant for $L = 2$.

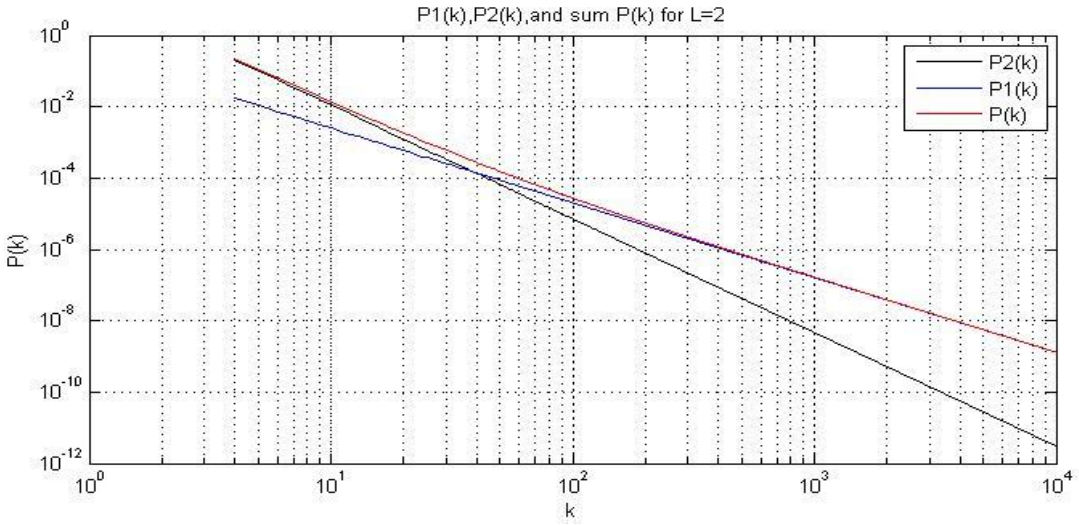


Figure 5.2. $P_1(k)$, $P_2(k)$, and sum of $P(k)$ for $L = 2$.

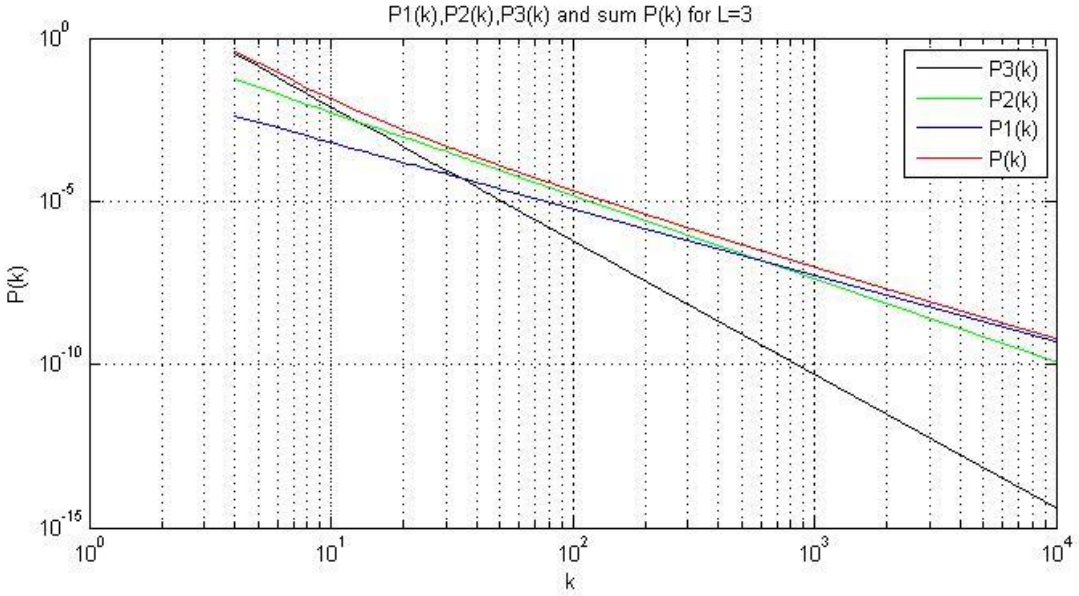


Figure 5.3. $P_1(k)$, $P_2(k)$, $P_3(k)$, and sum of $P(k)$ for $L = 3$.

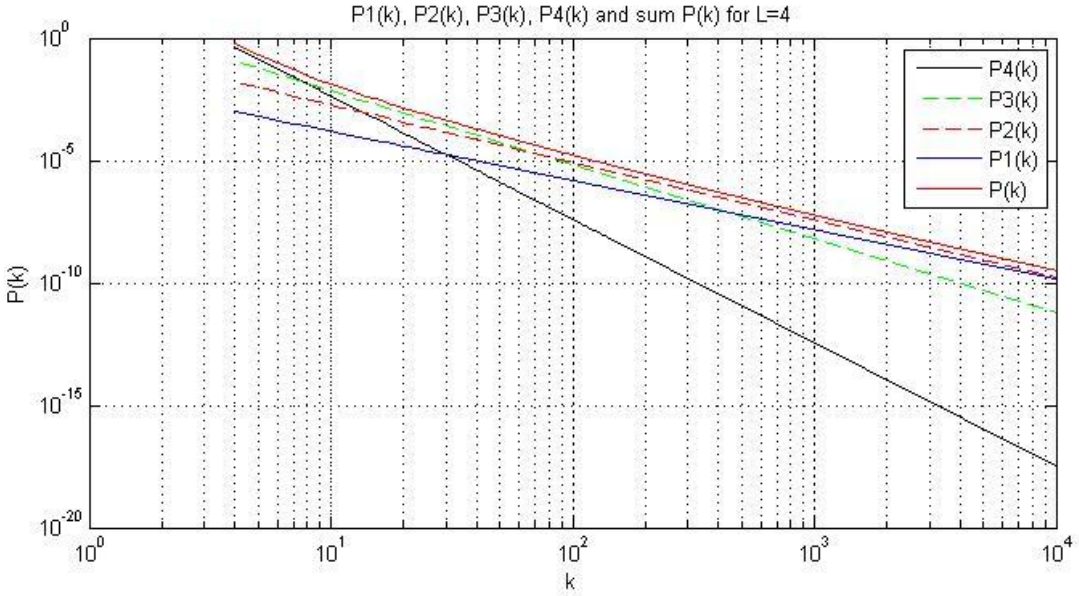


Figure 5.4. $P_1(k)$, $P_2(k)$, $P_3(k)$, $P_4(k)$, and sum of $P(k)$ for $L = 4$.

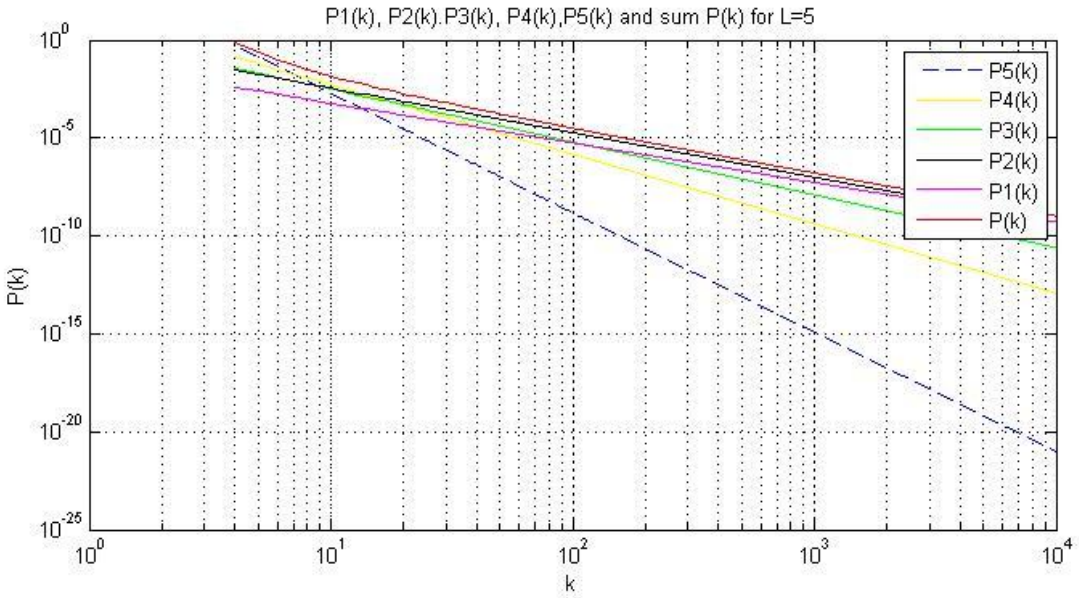


Figure 5.5. $P_1(k)$, $P_2(k)$, $P_3(k)$, $P_4(k)$, $P_5(k)$, and sum of $P(k)$ for L = 5.

The same procedure was repeated for L = 3, 4 and 5 as illustrated in Figures 5.3, 5.4 and 5.5. $P_i(k)$ is plotted for each of the PL exponent values, [γ_i , I =1, 2, 3,.., L], corresponding to the value of C that solves Equation 5.13 while satisfying the condition $C > (S_{ij})_{max}$= 1.0. Also, the sum, total $P(k)$, of each $P_i(k)$ is plotted. Comparison indicates that for large values of k, $P_i(k)$ for $\gamma_{min} = 1 + C$ is the closest to sum, $P(k)$ and can be considered the effective exponent in the summation.

Table 5.2 shows the accepted values of C satisfying the condition $C > (S_{ij})_{max}$ = 1.0 for different values of L and their corresponding values of PL exponents γ. Table 5.2 shows the different values of γ for different values of L and shows that for each value of L the node's degree distribution has PL exponent ranging from γ_{min} to γ_{max}.

Figure 5.6 is a plot of the accepted values of C, as shown in Table 5.2, and indicates that the value of accepted C is slowly and asymptotically approaching 1.0 as L increases, L >> 5.

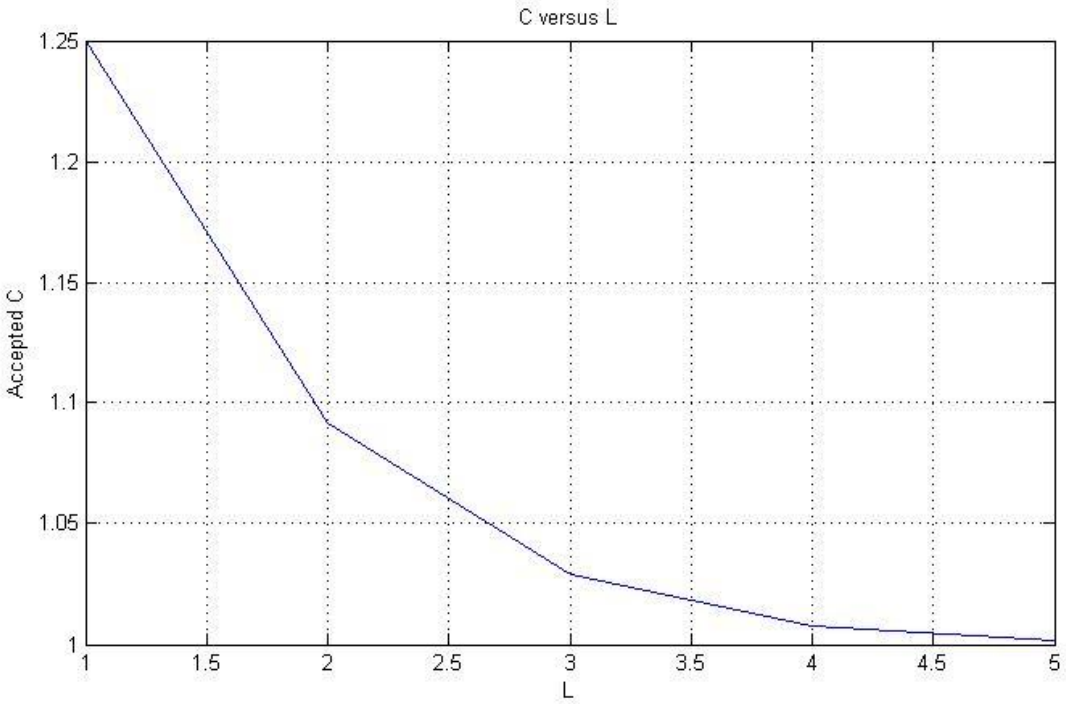


Figure 5.6. Accepted values of *C* versus *L*.

The value of γ_{max} corresponding to the accepted *C* values increases as *L* increases having value slightly larger than (*L*+1) as seen in Figure 5.7. The difference $[\gamma_{max} - (L + 1)]$ decreases as *L* increases, asymptotically approaching 0 as *L* becomes very large, *L* >> 5. Therefore, γ_{max} asymptotically approaches $(L + 1)$ as *L* becomes very large, *L* >> 5.

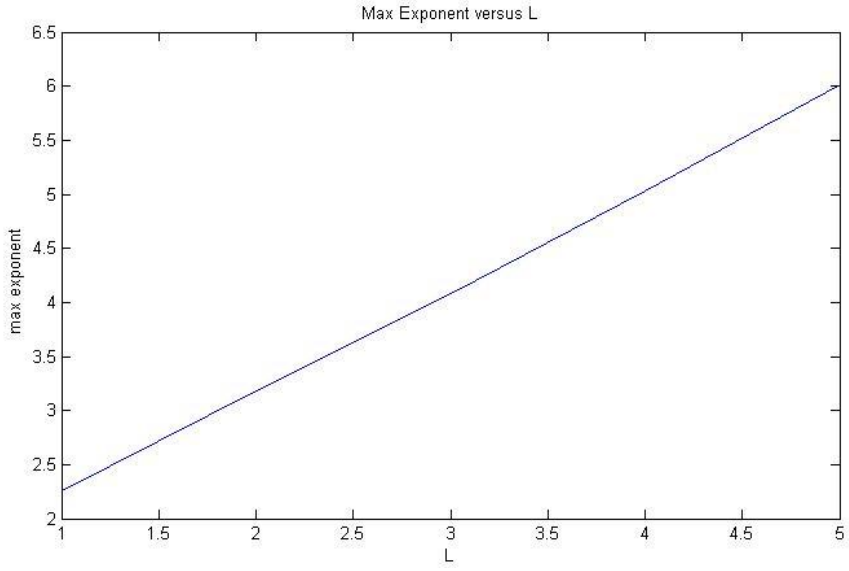


Figure 5.7. Values of γ_{max} versus *L*.

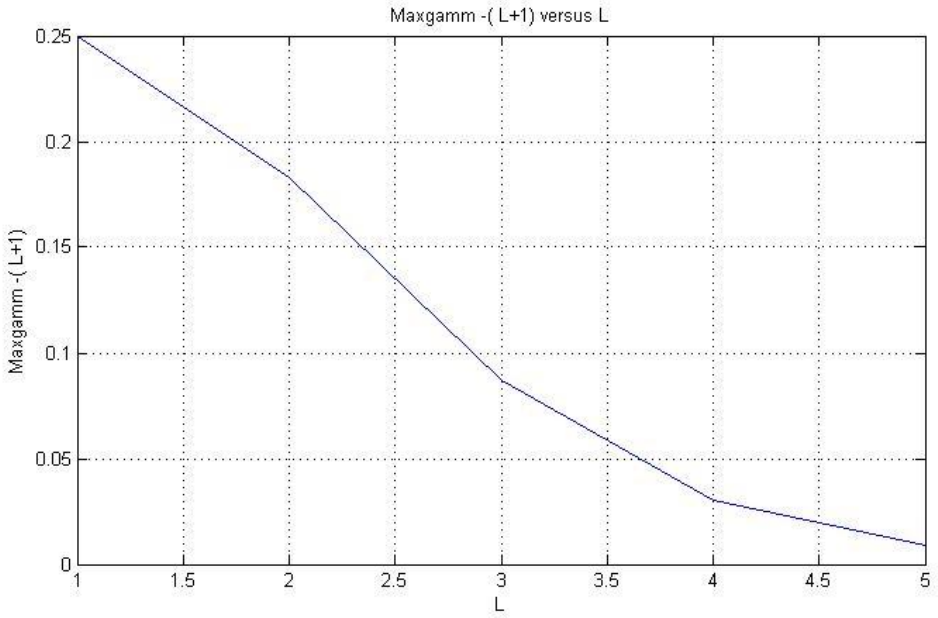


Figure 5.8. Values of $[\gamma_{max} - (\mathrm{L} + 1)]$ versus L.

On the other hand, Figure 5.9 indicates that the value of γ_{min} decreases as L increases and γ_{min} asymptotically approaches 2 as L becomes very large, L >> 5. Thus, for large values of L, γ_{max} approaches L+1 and γ_{min} approaches 2. Additionally, we deduce from Figures 5.2, 5.3, 5.4, and 5.5 that γ_{min} becomes the dominant exponent for large values of k.

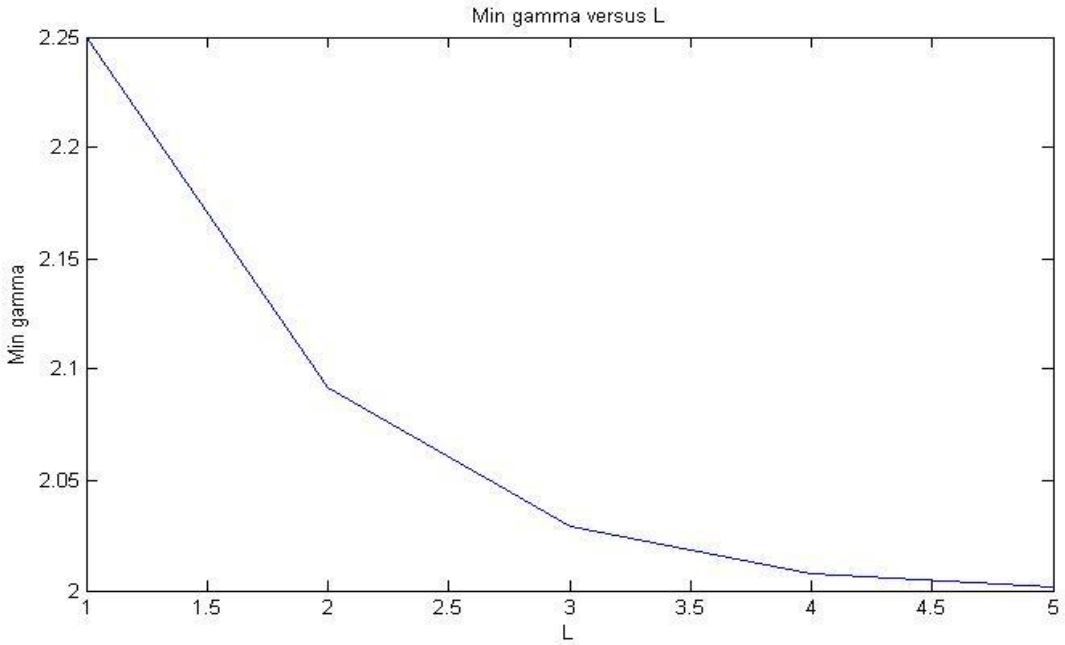


Figure 5.9. Values of γ_{min} versus L.

5.2.5. Discussion

The BA model produces a power law distribution given by $P[k] = 2(m)^2(k)^{-3}$. Thus, the BA model has a constant exponent, γ = 3 which is different from what is found in real-world complex networks where γ ranges from 1 to ∞ as stated in Chapter 2. For the case of IASM with a multiplicative attribute similarity measure, we analytically demonstrated that the IASM model produces networks with degree distribution given by:

$$P[k] = \sum_{w=1}^{L} \left\{ \frac{LC}{w} m^{(\frac{LC}{w})} F(w) \right\} k^{-((\frac{LC}{w})+1)} \quad , \gamma_w = -(1+C), -(1+C\frac{L}{L-1}), -(1+C\frac{L}{L-2}), \ldots, -(1+LC).$$

Thus, for length *L* node attribute vectors, the degree distribution of IASM is the sum of *L* power law degree distributions having exponents $\gamma_w = \left(\frac{LC}{w}\right) + 1$ where, *w* = {L,…, 3, 2, 1} for the accepted value of *C* given as the root of Equation 5.13 and $C >= 1.0$. Our numerical analysis of the distribution shows that at *L* = 1, the degree distribution is a PL with exponent $\gamma = 2.25$. As for the case where *L* = 2, it is the summation of two PL distributions. But, at higher values of *k*, the lower exponent dominates the higher exponent.

Hence, we conclude from this analysis that IASM is successful in preserving the power law degree distribution found in real-world complex networks. Additionally, IASM, unlike the BA model, reflects the heterogeneity of the nodes' attributes. Thus, IASM is capable of modeling the common scenario in real-world complex networks where a new node usually prefers to connect with existing nodes that are the most topologically popular and that have interests or attributes that are similar to those of the new node.

5.3. SNAM Analysis

5.3.1. Introduction

Simulation results that were previously presented in Chapter 4 show that SNAM preserves the power law degree distribution found in real-world complex networks. Networks generated using the SNAM model show high values of the average clustering coefficient similar to those of real-world networks. It was also shown that values of power law degree distribution exponent as well as the average clustering coefficient for networks generated using the SNAM model depend on many model parameters, especially *NoT* which represents the number of tests performed by each node to establish each of the *m* new links. This gives SNAM the capability to generate complex

networks with different statistical properties depending on the value of SNAM parameters used during network evolution. To verify the power law dependence found through simulation and to better understand SNAM, we analytically derive an expression for the nodes' degree distribution using the rate equation analysis method.

To make the analysis tractable, we need to make some modifications to SNAM as previously described. For convenience, we refer to the version of SNAM implemented and described previously as SNAM and we refer to the modified version of SNAM used for the analysis in this chapter as SNAM".

These modifications include the following.

1) In SNAM, the value of the number of tests, n_t or *NoT*, changes during the evolution of the network such that its value is incremented when reaching some predefined milestones during network evolution. Whereas in SNAM", n_t has a constant value during the entire network evolution. This assumption is made to facilitate the mathematical analysis to be able to reach a closed form equation for the degree distribution of the generated network. Having n_t incremented when reaching predefined milestones would not alter the power law degree distribution form of variation and its precise dependence on the value of n_t that is used. Thus, if the degree distribution power law dependence exists in SNAM", it would still exist in SNAM, but in a more complex form. In particular, the value of the overall power law exponent for SNAM depends on the initial value of n_t and the number of network growth milestones. In SNAM", the assumption that n_t is constant simplifies the power law exponent dependence to a dependence on a constant parameter n_t.

2) In SNAM, if node l reaches its maximum number of tests, *NoT*, and it still has not established its m connections, then arriving node l reduces its connection standard by a certain constant parameter ϵ. The testing of randomly chosen existing nodes is resumed using this new reduced connection standard for the same maximum number of tests n_t. This sequence is repeated until the arriving node l completes its m connections. In SNAM", we assume that the arriving node never has to decrease its connection standard. Thus, it is assumed that the probability of the arriving node failing to make all its m connections during n_t tests is negligible. Therefore, the present analysis assumes that the value of all arriving nodes connection standards are sufficiently low and that n_t is

sufficiently large such that the assumption that the new node will establish all its m connections during n_t tests per link is valid.

3) The SNAM connection standard, *CF*, depends on the normalized degree value and/or nodes attribute similarity. In this analytical solution, the SNAM" connection standard, *CF*, depends only on the normalized degree, similar to the BA model, for simplification of the mathematical analysis. Note that, here, we are not modifying the connection algorithm. Making the *CF* dependent on the normalized degree only simplifies the calculation method of the connection parameter *CF*, but the connection algorithm is not affected.

5.3.2. SNAM": Special Case of SNAM Considering Only Structural Popularity

The model starts with a seed network with m_o nodes connected by e_o edges given that $m_o \geq 2$ and $e_o \geq 1$.At each time step, a new node l is added to the network with m links to be connected to the existing network, one by one, where $m \leq m_0$. The connection function used here is as in the BA model. The connection of a new node l to old existing node i depends on the connectivity k_i of node i and is given by:

$$p_i = \frac{k_i}{\sum_j k_j} = p \tag{5.16}$$

In SNAM, each of the arriving new nodes has its connection standard S_l which describes its minimum requirement when making connections to existing nodes. We perform a test to see if the randomly chosen existing node connection function value exceeds or equals the connection standard of the arriving node. In the present analysis, existing nodes to be tested for connection with the new node are chosen randomly with replacement since analytically finding a closed form for the probability density function after excluding previously tested existing nodes would prove to be challenging or impossible.

This test is repeated n_t times for each of the new node's m links, which is the number of independent tests used to find suitable old existing nodes to be connected to the new node, where $2 \leq n_t \leq N_e$, where N_e is the number of existing nodes in the network.

The connection to a node can take place in any of consecutive tests.

1. Success on the first test, so a connection is established on the first test with probability p, or
2. Failure on the first test and success on the second, so a connection is established on the second test with probability $(1-p)p$, or
3. Failure on the first two tests and success on the third, so a connection is established on the third test with probability $(1-p)^2p$, and so on…… until failure on (n_t-1) tests and success on n_t, so a connection is established on test n_t with probability $(1-p)^{(n_t-1)}p$.

Since the tests are mutually exclusive, this connection or link from new node to an existing test node *i* during n_t tests is made with probability $\boldsymbol{P}$

$$\boldsymbol{P} = p + (1-p)p + (1-p)^2p + (1-p)^3p + \ldots + (1-p)^{(n_t-1)}p$$

$$= n_tp + f(p^2, p^3, \ldots) \quad (5.17)$$

Since, as in Equation 5.16, $p$is the normalized degree of existing nodes, $p \ll 1$, the higher orders of p terms of Equation 5.17 can be neglected w.r.t. the value of n_tp. Therefore, $\boldsymbol{P} \cong n_tp$.

An alternate proof for expression of $\boldsymbol{P}$i s:

P[successful connection within n_t tests] = 1 - P[no connection within n_t tests]

Since P[no connection within n_t tests] = P[no connection within first test] and P[no connection within second test] and P[no connection within third test] and P[no connection within fourth test] … and P[no connection on n_tth test].

Since each (no connection) in a test is statistically independent of (no connection) in other tests.

P[no connection within n_t tests] = $\prod_{l=1}^{n_t}$ P[no connection on l^{th}trial]

$$= \prod_{l=1}^{n_t}(1-p)^l \quad = \quad (1-p)^{n_t}$$

Therefore, $\mathbf{P} =$ P[successful connection within n_t tests]

$$= 1 - (1-p)^{n_t} = n_tp + f(p^2, p^3, \ldots) \cong n_tp \quad (5.18)$$

Equations 5.17 and 5.18 are the same.

Additionally, as explained above, we take small enough values of the node connection standards and high enough values of n_t so that it can be assumed that a connection will be made during the n_t tests.

Since this process is repeated for each of the m links to be established per unit step, the rate equation representing the rate at which the node i acquires edges can be written as follows.

$$\frac{\partial k_i(t)}{\partial t} = m[n_t p] = m\left[n_t \frac{k_i}{\sum_j k_j}\right] = \frac{m\, n_t k_i(t)}{2e_o + 2mt}$$

Therefore, for large values of t,

$$\frac{\partial k_i(t)}{k_i(t)} = \frac{n_t}{2}\frac{\partial t}{t} \tag{5.19}$$

Integrating both sides of Equation 5.19 gives

$\ln k_i(t) = \ln(t)^{\frac{n_t}{2}} + c$, where c is a constant.

Since $k_i(t_i) = m$, since t_i is the time at which node i was added to the system.

$\ln m = \ln(t_i)^{\frac{n_t}{2}} + c$, where c is a constant.

Therefore, $\ln \frac{k_i(t)}{m} = \ln(\frac{t}{t_i})^{\frac{n_t}{2}}$

or $k_i(t) = m(\frac{t}{t_i})^{\frac{n_t}{2}}$ (5.20)

The cdf of $k_i(t) =$ the probability that a node i has a connectivity smaller than k

$$= P[k_i(t)<k] = P\left[m\left(\frac{t}{t_i}\right)^{\frac{n_t}{2}} < k\right] = P[\mathrm{t}_i > \left(\frac{m}{k}\right)^{\frac{2}{n_t}} \mathrm{t}]$$

Therefore, $P[k_i(t)<k] = 1 - P[t_i \leq \left(\frac{m}{k}\right)^{\frac{2}{n_t}} t]$ (5.21)

Since the initial seed size is m_o and one new node is added uniformly at each time step,

$$P(t_i) = \frac{1}{m_o + t}$$

Therefore, $\mathrm{P}[k_i(t)<k] = 1 - \frac{(\frac{m}{k})^{\frac{2}{n_t}} t}{m_o + t}$

The probability density of a node having degree *k*, P[k] can be obtained using

$P[k] = \partial \mathrm{P}[k_i(t) < k]/\partial k$ giving

$$P[k] = \frac{\partial \mathrm{P}[k_i(t)<k]}{\partial k} = \frac{2}{n_t}\frac{(m)^{\frac{2}{n_t}}t}{m_o+t}(k)^{-(\frac{2}{n_t}+1)}$$

Since at long times $t >>> m_o$, $m_o + t \cong t$ and

$$P[k] = \frac{2}{n_t}\frac{(m)^{\frac{2}{n_t}}t}{t}(k)^{-(\frac{2}{n_t}+1)} = \frac{2}{n_t}(m)^{\frac{2}{n_t}}(k)^{-\gamma}$$

Therefore, $P[k]$ follows a power law distribution with exponent $\gamma = \frac{2}{n_t} + 1$ for $n_t \geq 2$.

5.3.3. Discussion

We note that the BA model has produced a power law distribution given by $P[k] = 2(m)^2(k)^{-3}$. Thus, the BA model has a constant exponent, γ, of value 3 which is different from what is found in real-world complex networks where γ ranges from 1 to ∞ as stated before in Chapter 2. SNAM (actually SNAM") produced a network with power law distribution given by $P[k] = \frac{2}{n_t}(m)^{\frac{2}{n_t}}(k)^{-\gamma}$, where $\gamma = \frac{2}{n_t} + 1$ for $n_t \geq 2$. This gives SNAM an advantage over BA as SNAM is capable of generating complex networks with adjustable statistical properties. In SNAM, changing the number of tests, n_t, yields different types of complex networks with different statistical properties.

Hence, we conclude from this analysis that SNAM is successful in preserving the power law degree distribution found in real-world complex networks and that the value of the power law exponent depends on the value of n_t used in network generation. SNAM introduces a new concept of node heterogeneity which is the heterogeneity of the nodes' requirements to establish a connection. SNAM, as far as we know, is the only model for complex network generation that considers individual differences between nodes by assigning this heterogeneous connection standard parameter to the arriving nodes. Additionally, SNAM excels in its capability of generating specific types of complex networks by varying the model parameters.

5.4. Validation

To provide some validation of our results, we compare the results obtained from the mathematical analysis of our models, as discussed above, with earlier simulation results, as presented in Chapter 4. We generate networks using our MATLAB simulation code for the special cases of IASM_A and SNAM that were analyzed in our mathematical analysis. We try to incorporate the assumptions made during the mathematical analysis in our simulations. However, this can prove to be challenging or not applicable as in theoretical analysis where everything can be assumed ideal and usually differs from reality.

Using IASM_A, networks with 1,000 nodes are generated via MATLAB with $m = m_o = 5$, $L = 10$, $\alpha = 1$, $w = 0$, and $\beta = 0$ (multiplicative degree and attribute similarity *CF*). The networks are generated for different attribute vector lengths, *L*. The PL exponent magnitude values of these simulations are plotted together with those resulting from the mathematical analysis of IASM_A, previously shown in Figure 5.9. From comparison of the two PL exponent curves in Figure 5.10, we can see that the analytical and simulation results are similar in that the simulation results are very close to the analytical results and both curves have the same tendency of magnitude decreasing as the number of attributes, *L*, increases.

Additionally, we use SNAM to generate networks with 1,000 nodes using our MATLAB simulation code with $m = m_o = 5$, $L = 10$, $\alpha = w = 0$, and $\beta = 1$ (degree only CF). The networks are simulated for various values of €. Figure 5.11 shows the plot for PL exponent magnitude values resulting from both the analytical analysis and simulation (with variable €) versus the *NoT* parameter. The figure shows that the PL exponent magnitude values from the analytical analysis are close to values resulting from simulation and follow the same behavior when the *NoT* parameter is varied.

5.5. Conclusion

In this chapter, we have used the rate equations for models IASM and SNAM along with mean field theory to find expressions of the degree distribution for each model. We concluded that for IASM having the connection function, *CF*, dependent on the structural popularity multiplied by the attribute similarity, the degree distribution is a sum of PL degree distributions having exponents $\gamma_w = \left(\frac{LC}{w}\right) + 1$ where, $w = \{L,\ldots, 3, 2, 1\}$. Additionally, SNAM" whose *CF* is

dependent only on structural popularity, has a power law degree distribution given by $P[k] = \frac{2}{n_t}(m)^{\frac{2}{n_t}}(k)^{-\gamma}$, where $\gamma = \frac{2}{n_t} + 1$ for $n_t \geq 2$.

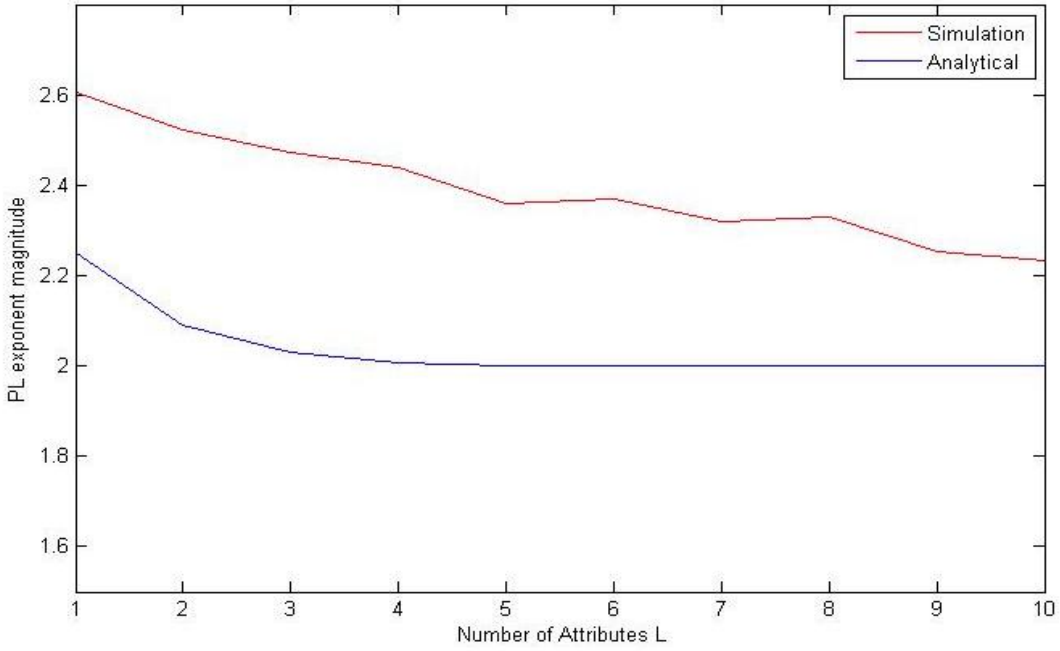


Figure 5.10. Power law exponent magnitude versus number of attributes, *L*, derived by analytical and simulation methods for IASM_A.

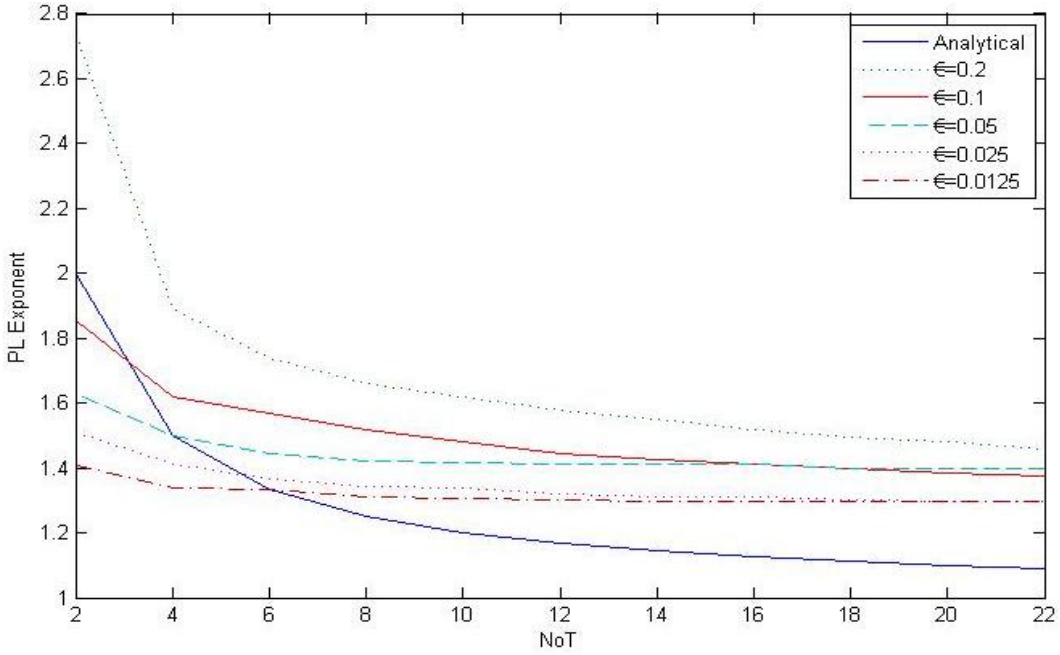


Figure 5.11. Power law exponent magnitude versus number of trials, *NoT*, derived by analytical and simulation methods for SNAM.

Chapter 6. A Case Study Using the Heterogeneous Complex Network Generation Models

6.1. Introduction

This chapter discusses evaluating our proposed models, Integrated Attribute Similarity Models (IASM) and Settling Node Adaptive Model (SNAM), using an online social network modeling case study. The reasons for choosing an online social networks (OSN) for the case study are first discussed. Examples of OSN applications in different fields are presented next. The mathematical modeling of OSNs is then discussed. Next, we introduce the benefits and suitability of mathematically modeling OSNs using our models. The method for evaluating IASM and SNAM using the case study is presented. Evaluation results are then presented. Finally, we discuss the conclusions of the case study.

6.2. Using Online Social Networks for the Case Study

Our proposed models are intended to be general and, hence, can be used to generate different types of complex networks used in different application domains. However, as a proof of concept and to provide a more in-depth study of the application of our models to one particular type of complex network, we conduct a case study where we apply the models developed in this research to online social networks.

Sociologists have long used the term social networks (SNs) to represent interconnections, between different entities that are formed for reasons such as similar interests or context (location or job). An online social network (OSN), which is of interest here, is defined as a digital representation of the relations between registered entities, individuals or institutions [2]. Our choice of online social networks is mainly due to their widespread use in a large number of application areas, including marketing, information diffusion, recommendation, and trust analysis [43]. Online social networking websites are used to maintain, strengthen, and support offline social relations. OSNs contain within them a lot of information, such as the relations between actors or registered entities (topological structure) and semantic information about the actors or their published content. The importance of OSNs has led to their being studied widely [44] and makes them a good candidate for a case study for this research.

6.2.1. Applications of Online Social Networks

Data extracted from OSNs is used to guide decision makers to make more accurate decisions in applications such as healthcare, marketing, and information diffusion [43]. Additionally, OSNs can be used for answering numerous research questions in the application domain, such as: What are the different communities in the network and their members? What customer(s) would most likely be interested in a certain product? Who are the most influential users? What is the susceptibility of users to disease?

6.2.2. Mathematical Modeling of Online Social Networks

A good mathematical model for OSNs should mimic the structure and dynamics of the actual OSNs. Thus, applying our models for generating an OSN should preserve the statistical properties exhibited by the real OSN, calculated or determined for this OSN by other researchers. As stated previously, OSNs are characterized by having a power law degree distribution, high average clustering coefficient, small average path lengths, and the emergence of community structure. Moreover, the models should be capable of accurately predicting future connections between network nodes or users.

6.2.3. Using IASM and SNAM to Generate Online Social Networks

We conjecture that IASM and SNAM will be useful in analyzing online social networks by generating networks that mimic the real network structure and dynamics. OSNs grow as users with similar interests create more connections with each other rather than with users with different interests. The set of node attributes is defined based on the characteristics or profiles of the social network users. Individual differences between users have an effect on their connections. Additionally, users in social networks usually make connections based on the structural popularity of other users. The connection standard of SNAM can be interpreted by individual differences between users making connections. To our knowledge, integration of structural popularity with multi-attribute similarity and including individual differences when making connections has not been represented in any other model prior to our research.

6.3. Dataset for the Case Study

A real dataset representing a real OSN will be used in the case study. This dataset should provide empirical results gathered about the OSN being studied. We opted for using a suitable dataset among datasets published through previous projects. We think that creating our own dataset is time consuming and does not add to proving the validity of our models. Performing the case study requires a true understanding of the dataset that is used. Thus, the statistical properties for the dataset must be obtained.

We chose one of the datasets for online social networks made available online by Stanford University, ego-Facebook [45]. The dataset consists of "friend lists" from Facebook and was collected using the Facebook application. The dataset chosen was for anonymous Facebook user connections. The dataset includes node attributes (profiles) and their undirected structural connections. The dataset consists of 4,039 users or nodes with 88,234 edges among those nodes. The dataset contains 10 ego users' networks. An ego user (node) is called the focal node and the nodes directly connected to it are commonly known as alters. Each of the 10 ego networks contains ties between its focal node and its connected alter nodes and the present ties between these alters. To perform statistical analysis on the network, we combined the ego networks, including the ego focal nodes themselves.

6.3.1. Processing the Dataset

We started by combining the data available from the 10 ego networks to obtain the adjacency matrix describing all network connections. Additionally, the node features were recorded in relation to the ego users so we had to extract the real node features values. After performing this analysis, we had a 4,039 × 4,039 square adjacency matrix of user connections with 1,283 feature (attribute) vector for each user. Thus, now each node has a binary vector of length 1,283 associated with it whose elements take either the value of 1 if the node possess this attribute or the value 0 if the node is not interested in this attribute.

We calculated the statistical properties for the extracted adjacency matrix. The value of the power law exponent for the adjacency matrix was found to be -1.1697, the average path length was 3.6925, and the value of the average clustering coefficient was 0.6055. To validate our network growth models using real-world network information as a case study, a small subset of

the real-world network connections was used as a seed for synthetic network growth. A MATLAB program used the seed adjacency matrix structural information, together with attributes of all nodes of the real network, to grow a synthetic network with the mathematical model to be validated. The synthetic network has the same size as the real network size. Then, statistical properties of the real and synthetic networks were compared. Study of the emergence of the community structure property was not applicable as the dataset did not contain information about the network communities.

6.4. Evaluation of IASM and SNAM Using the Facebook Dataset

We believe that our model will be useful in the field of link prediction for social networks, which has many applications in fields such as marketing, terrorist network analysis, and healthcare. Thus, to validate and evaluate the accuracy of our model's link prediction, we introduce the attribute vectors for the nodes and a subset of the adjacency matrix as model inputs. The objective is to use the seed network extracted from the real network dataset, together with the whole real network attribute data, to generate an OSN of the same size as the studied network dataset. Network evolution is done using our IASM or SNAM models. Additionally, we generate networks of the same size using the Erdös and Rényi (ER) model. Generating a network of the same size and number of edges with the ER model gave different statistical properties from that of the dataset. The network generated with the ER model has an average path length of 2.198 and an average clustering coefficient of 0.010819. The statistical properties of the real network of the dataset will be compared to those of the networks generated using our models via simulation by MATLAB. Additionally, we compare the network generated using the BA model with the Facebook dataset network which is presented in the next section as a special case of using our IASM_A model having CF parameters $\alpha = 0$, $\beta = 1$, and $w = 0$.

6.4.1. Generating the Facebook Network via IASM_A

Next, we generate networks of the same size as the real dataset network using IASM_A with different combinations of *CF* coefficients values. The attribute vectors of all 4,039 Facebook dataset nodes is an input to our IASM_A generation model. Since the order of the time arrival of the different nodes is unavailable, it is assumed that the node Id in the dataset's adjacency matrix is its order of arrival. A subset of the Facebook dataset adjacency matrix, having 100 nodes, is

used as the seed network for IASM_A. Networks are generated for different *CF* coefficients values for both cases of the model with and without a triad formation step (TFS).

6.4.1.1. Simulation Results

IASM_A, when used to generate networks of the same size, grows networks having the statistical properties shown in Table 6.1. As mentioned in Section 6.4 the IASM_A model without TFS reduces to the BA model when $\alpha = 0$, $\beta = 1$, and $w = 0$.

Table 6.1. Simulation Results for Networks Generated with IASM for Different *CF* Coefficients

			Without TFS			**With TFS=1**		
α	***w***	***β***	**Exp_PL**	**Av_Pl**	**Av_CC**	**Exp_PL**	**Av_Pl**	**Av_CC**
0.9	0	0.1	-1.7988	2.6165	0.0172	-2.1484	2.7586	0.2214
0	0.5	0.5	-1.6770	2.6201	0.0149	-2.294	2.7733	0.2289
1	0	0	-1.7620	2.6183	0.0169	-2.2040	2.7527	0.2213
0	0	1	-2.0349	2.6052	0.0174	-2.4147	2.7346	0.2040

6.4.1.2. Analysis of Simulation Results

Even though IASM_A did not produce the same power law exponent found in the Facebook dataset (-1.1697),the percentage error in obtaining the power law exponent is reduced from about 74 percent in the case of using the BA model to about 50 percent, 43 percent, and 53 percent when using IASM_A with different combinations of CF coefficients values [$\alpha = 1$, $w = \beta = 0$], [$\alpha = 0$, $w = \beta = 0.5$] and [$\alpha = 0.9$, $\beta = 0.1$, $w = 0$], respectively. These are the parameters shown in Table 6.1. However, both IASM and the original BA produce a shorter average path than that found in the Facebook dataset. Additionally, both do not produce the high average clustering coefficient found in the real Facebook dataset. Adding a TFS increased the clustering coefficient, but it also increased the power law exponent magnitude and slightly increased the average path length.

6.4.2. Generating the Facebook Network via SNAM

This section describes the objective of using SNAM to generate a network of size equal to that of the Facebook dataset. The SNAM algorithm implemented in MATLAB should use whole attribute vector corresponding to all 4,039 Facebook nodes as an input along with a sample subset of Facebook nodes which act as seed network with $m_o = 100$ nodes for SNAM.

The original SNAM model presented in Chapter 3 and analyzed in Chapters 4 and 5was developed to be able to grow complex social networks characterized with certain ranges for the power law exponent, average path length, and average clustering coefficient. The dataset for the Facebook network that we used for the case study is a bit different since it is characterized by a power law exponent having magnitude much lower than the magnitudes cited in publications for complex social network statistical characteristics. Nevertheless, we were able to introduce some changes in the parameters as well as the steps governing the growth process of the synthesized Facebook network to produce results that approach the statistical characteristics of the real Facebook dataset used for the case study.

6.4.2.1. Simulation Results with Original SNAM

The first trials to generate the synthetic Facebook network using the original SNAM algorithm steps resulted in the failure of some nodes to connect to *m* pre-existing nodes upon birth. Thus, we were faced with the problem of unconnected nodes and the algorithm running indefinitely.

As stated in previous chapters, the SNAM model is based on allowing the new added node to perform multiple tests with existing nodes and to make a connection to an existing node that meets the new node's standard. If the new node reach its maximum number of tests, *NoT*, and it still does not make its *m* connections, then the new node must lower its standard. The new node after lowering its standard continues testing old nodes until it either makes its *m* connections or it reaches its maximum value for *NoT* again. The new node can continue reducing its standard indefinitely until it makes its *m* connections. This SNAM algorithm worked well assuming that any new node will have some similarities with old nodes as a result of using node attribute vectors whose binary elements uniformly distributed. Thus, no new node is expected to have null attribute similarity values with all old nodes.

6.4.2.2. Enforcing Reduction Limit *R*

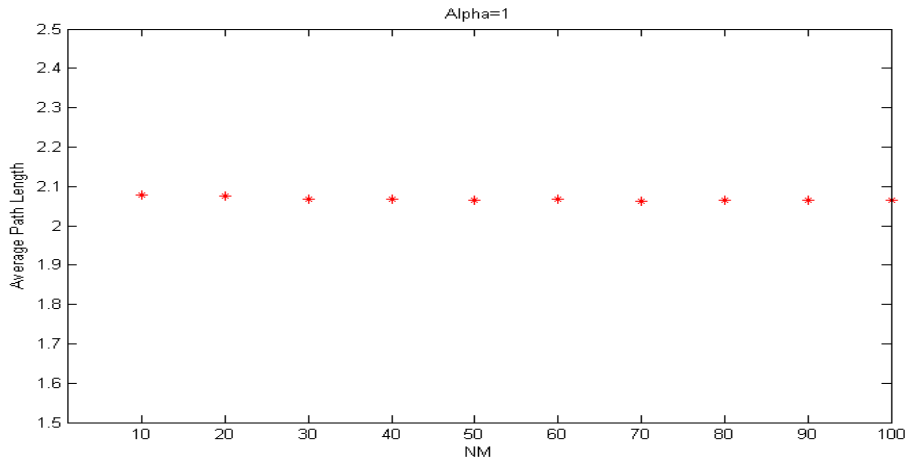
Given the results above, we further investigated the characteristics of the Facebook dataset. We observed that some network nodes in this dataset possessed unique attribute values causing them to have low or null attribute similarity values with other network nodes. This is expected to cause these nodes to either have no connections or fail to complete their *m* connections. The original SNAM algorithm will spend a lot of time when any such new node arrives as it searches for *m*

old nodes that are somewhat similar to it and continues to reduce its connection standard *S* indefinitely.

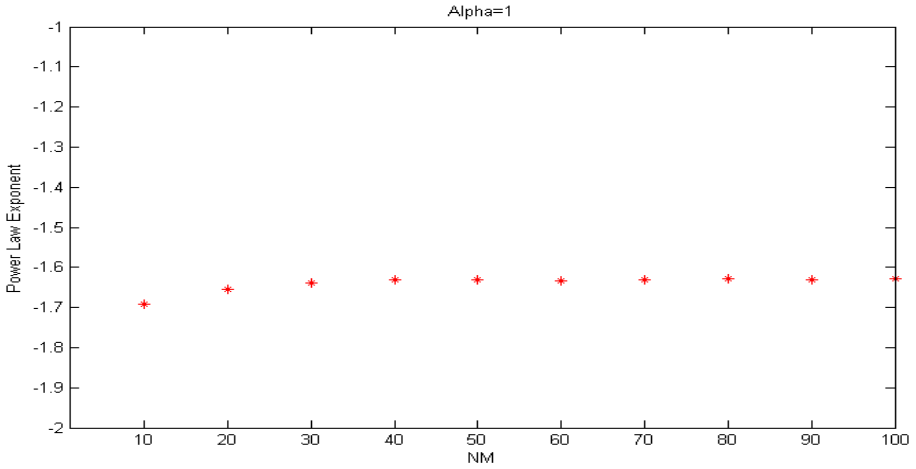
Therefore, we introduced modifications to our original SNAM model to approach the unique characteristics of the particular dataset used for the case study. We noticed that a new added node that has a low *CF* value with old network nodes is equivalent to this new node having a high connection standard (high value of *S*). Thus, we introduced a new model parameter *R* that puts a limit on the number of reductions on the value of *S* by €. Therefore, *R* puts a constraint on the extent of decreasing a new node connection standard until it settles for an old node whose *CF* value is less than the new node's reduced standard. The new node with a high standard starts searching for nodes that meet its standard. After reaching its maximum number of tests, *NoT*, it starts reducing its *S* value. The original SNAM allows the value of *S* to be reduced indefinitely until a connection is made and this is repeated until it makes *m* connections. However, now the new node reduces its *S* value by € only *R* times. Therefore, such new nodes after making all *R* reduction on the new node's value of *S* may have fewer than *m* connections or may even end up with no connections. A node ending this process with fewer than *m* connections randomly selects nodes to which it connects. These random connections ensure that there are no unconnected nodes and that each new node has a minimum of *m* links to existing nodes. We denote this modified version of SNAM as SNAM*.

The SNAM* connection algorithm allows both types of nodes (having zero or fewer than *m* connections after exhausting *R* standard reduction steps) to complete their *m* links either through the model or through random completion of connections. The set of results for this first modification to SNAM are shown in Figures 6.1, 6.2, and 6.3. The three figures show average path length, PL exponent and average clustering coefficients for *CF* coefficient values of [$\alpha = 1, w = \beta = 0$], [$\alpha = 0, w = \beta = 0.5$] and [$\alpha = 0.9, \beta = 0.1, w = 0$]. These are the sets of parameter values specified in Table 6.1 for Facebook growth using IASM_A. This use of such different *CF* coefficient combinations shows how the generated network statistical properties are affected when using *CF* coefficients based on using added or multiplied attribute similarity along with the degree centrality. The statistical properties are plotted with x-axis *NM*. *NM* was described in Chapter 4 as the number of milestones during network evolution at which the value of *NoT* is incremented by one. The average degree of the nodes in the generated network (*m* =

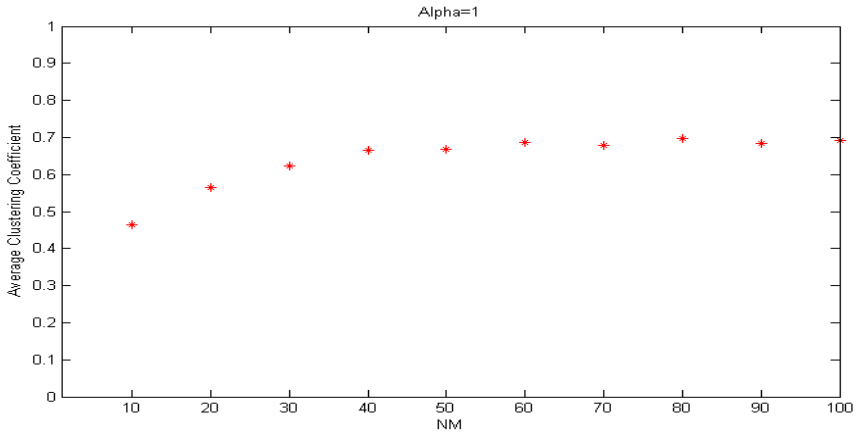
22) was approximated by dividing the number of the links in the original dataset by the final number of nodes in the network. The threshold € was given the value of 0.2 and reduction limit *R* = 24 was found by generating networks of smaller sizes and observing which *R* results in the growth of networks with high



(a) Average Path Length

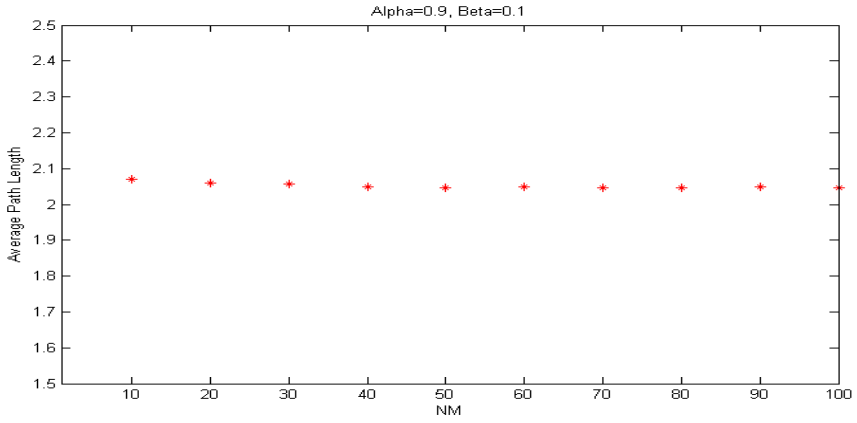


(b) Power Law Exponent

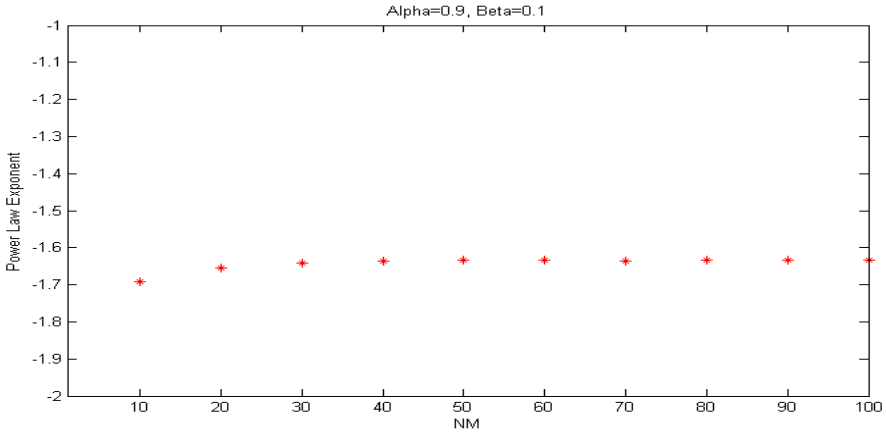


(c) Average Clustering Coefficients

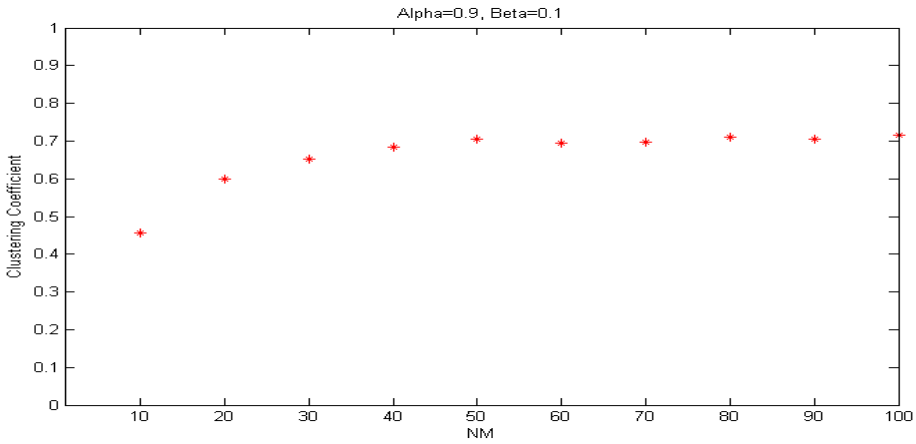
Figure 6.1. SNAM* algorithm with a normalized degree with multiplied attribute similarity *CF* ($\alpha = 1$, $w = \beta = 0$): a) Average Path Length, b) Power Law Exponent, c) Average Clustering Coefficients



(a) Average Path Length

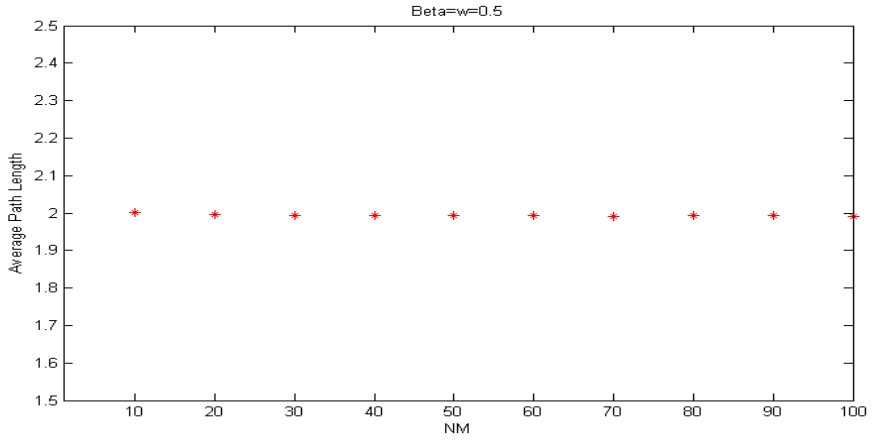


(b) Power Law Exponent

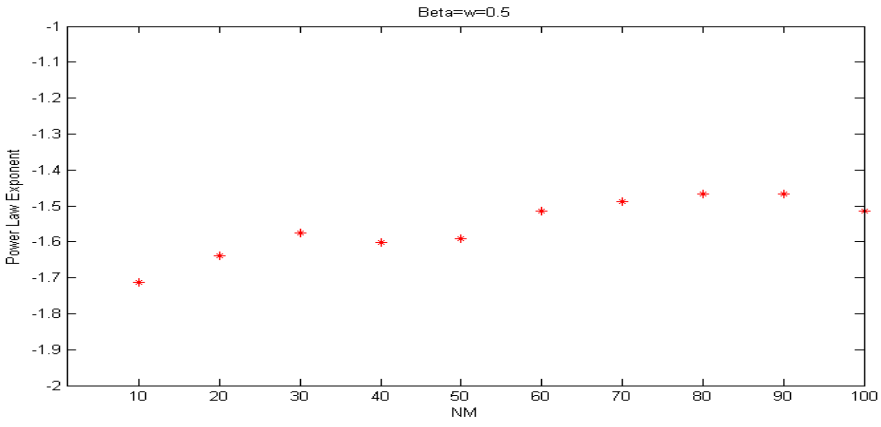


(c) Average Clustering Coefficients

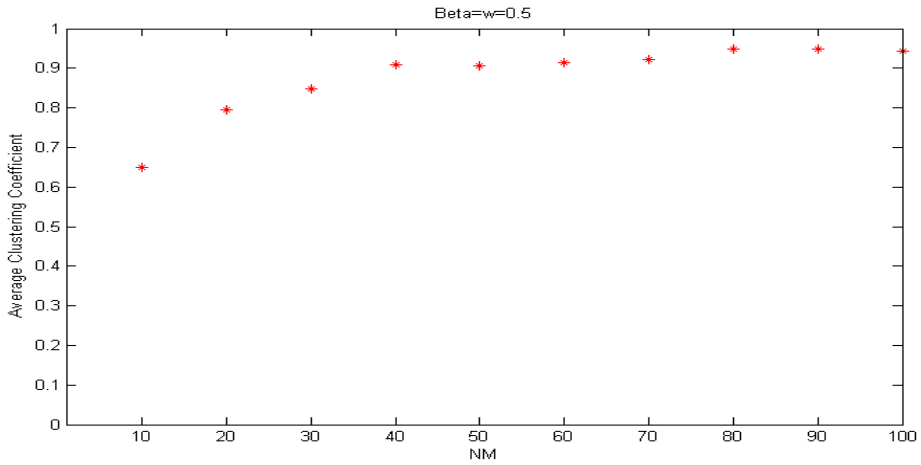
Figure 6.2. SNAM* algorithm with *CF* coefficients (α = 0.9, β = 0.1, *w* = 0): a) Average Path Length, b) Power Law Exponent, c) Average Clustering Coefficients



(a) Average Path Length



(b) Power Law Exponent



(c)Average Clustering Coefficients

Figure 6.3. SNAM* algorithm with a normalized degree with added attribute similarity *CF* ($\alpha = 0$, $w = \beta = 0.5$): a) Average Path Length, b) Power Law Exponent, c) Average Clustering Coefficients

average clustering coefficient values and that have power law exponent magnitudes lower than 2 as found in the original dataset.

6.4.3. Generating the Facebook Network with Extended SNAM

The results presented in Section 6.4.2 for SNAM* having reduction limit *R* were found to possess a statistical power law exponent and an average path length with values that differ significantly from those for the actual Facebook network dataset. Studying the Facebook dataset adjacency matrix led to the observation that the Facebook dataset has many nodes with only one or two connections. This is different from networks generated by our SNAM and SNAM* models where nodes have a minimum degree of value *m* which has a value greater than two here to generate a network with number of links near that of the original dataset.

A node with a high connection standard is expected to have fewer connections than other network nodes because of its high standards. But, in SNAM the connection standard *S* is reduced until all *m* connections to existing nodes are made. And, in SNAM* with the reduction limit described in Section 6.4.2, a node will randomly complete its *m* connections to existing nodes if they are not completed in the normal manner. Given the observation from the real Facebook dataset used in the case study that there are nodes in the real network with low degree

centrality values (0, 1, 2, …), we further modified the connection algorithm in SNAM* so that a new node with fewer than *m* connections, including zero connection, is linked to only one additional random node after exhausting its *R* reductions. This replaces the approach of continuously reducing the new node's *S* value until it makes *m* connections or completing the *m* connections randomly. We denote this further extended SNAM as SNAM**.

6.4.3.1. Simulation Results after Adding One Link when the Reduction Limit *R* is Reached

The following results show the statistical properties of the networks grown using SNAM**, forcing the nodes upon exhausting their *R* reductions to make only one additional connection. The results for the average path length, PL exponent and average clustering coefficients values in Figures 6.4, 6.5, and 6.6 are for the same combinations of *CF* coefficient values as used for Figures 6.1, 6.2, and 6.3.The x-axis variable *NM*, the number of milestones during network evolution, is varied from 10 to 100. The number of connections to each new born node *m*, was increased to $m = 60$ to compensate for the case where some nodes make less than *m* connections as a result of their making only one additional connection after exhausting their *R* reductions. The threshold € is given the value of 0.2 as before and $R = 16$ was found by generating smaller sized networks and finding the *R* value corresponding to the networks with statistical properties closest to that of the dataset.

6.4.3.2. Analysis of Simulation Results

From Figures 6.4 through 6.6, we can see that the power law exponents of the generated networks approach the values for the Facebook dataset used in the case study. Additionally, there is an increase in the average path for the generated networks although it is still less than that of the original dataset. Thus, observing the characteristics of the dataset's adjacency matrix and attribute vector and reflecting these characteristics in the modified SNAM algorithm, SNAM**, makes SNAM more capable of closely representing the particular Facebook dataset used in the case study.

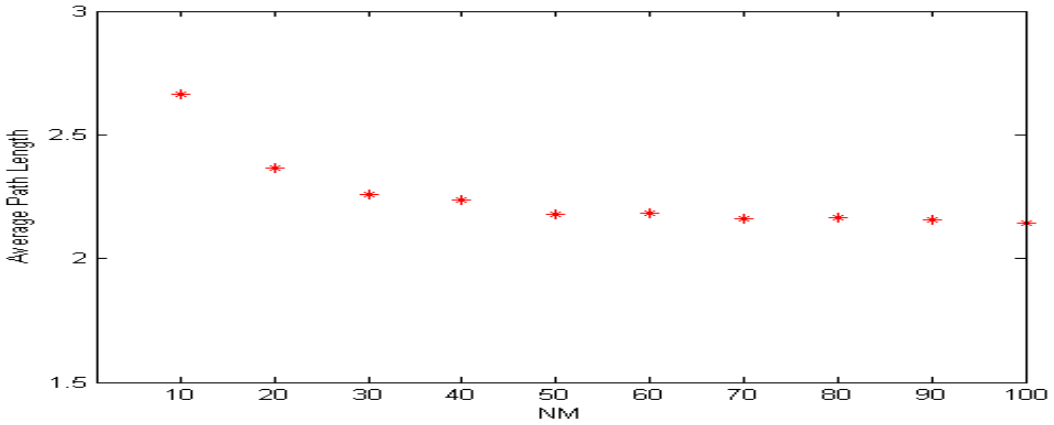
6.4.4. Effect of Reduction Limit *R* on Network Statistical Properties

The appropriate *R* value for the Facebook dataset was found by trial and error by generating networks of a smaller size than the dataset and observing the effect of different *R* values on the statistical properties of the generated networks.

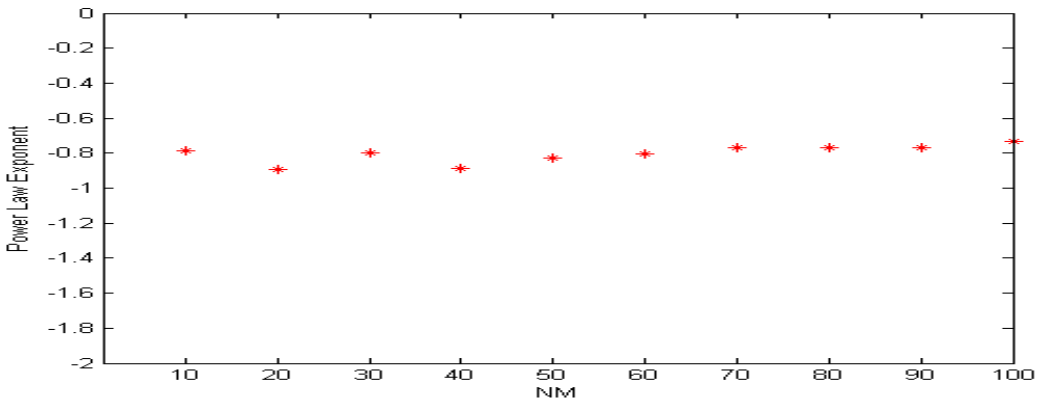
6.4.4.1. Simulation Results for the Effect of the Reduction Limit *R*

The results presented in this section show the effect that different *R* values have on the statistical properties of the generated networks. Networks are generated for the same *NM*, *m*, €, and *CF* coefficient values, but different *R* values. This was done for the version of SNAM that completes its *m* connections after exhausting its *R* reductions, denoted as SNAM* and the one that makes only one additional connection, denoted as SNAM**. The simulation parameters are *m* = 22 and € = 0.2 and *CF* coefficients α = 0.5, *w* = 0.5, and β = 0.

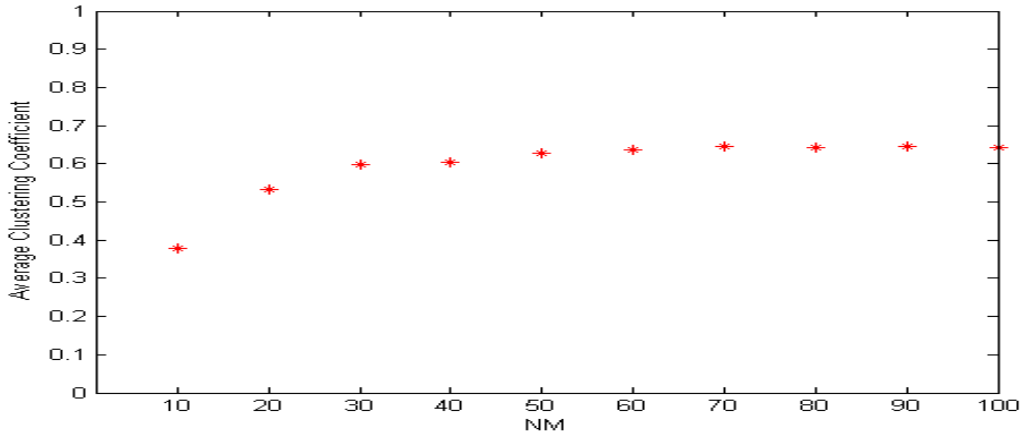
Tables 6.2 and 6.3 show the effect on the statistical properties of *R* taking different values ranging from 5 to 100. Values for Exp_Pl (power law exponent), Av_Pl (average path length), and Av_CC (average clustering coefficient) corresponding to different *R* values are presented in Table 6.2 for SNAM* and in Table 6.3 for SNAM**.



(a) Average Path Length

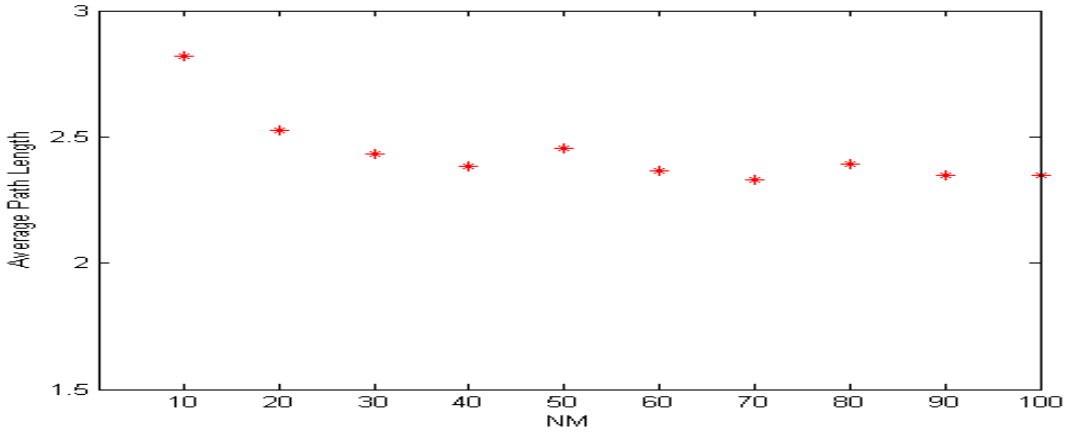


(b) Power Law Exponent

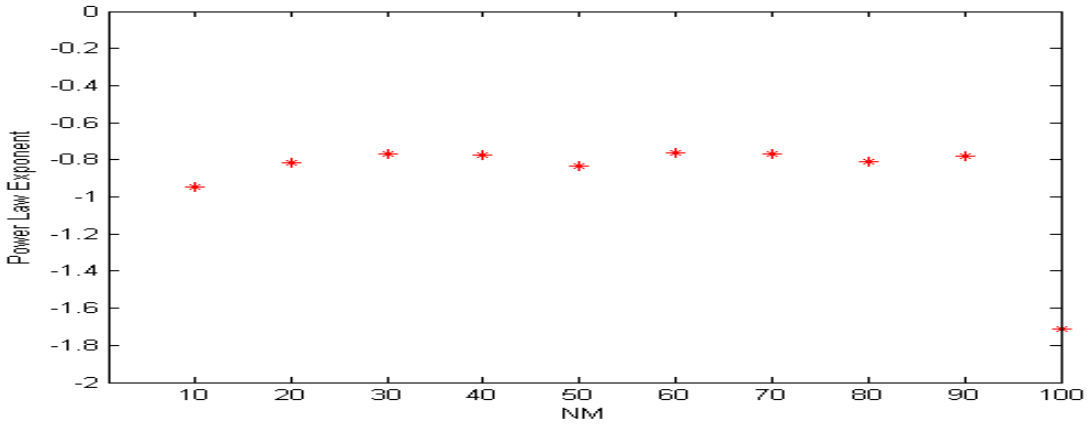


(c) Average Clustering Coefficients

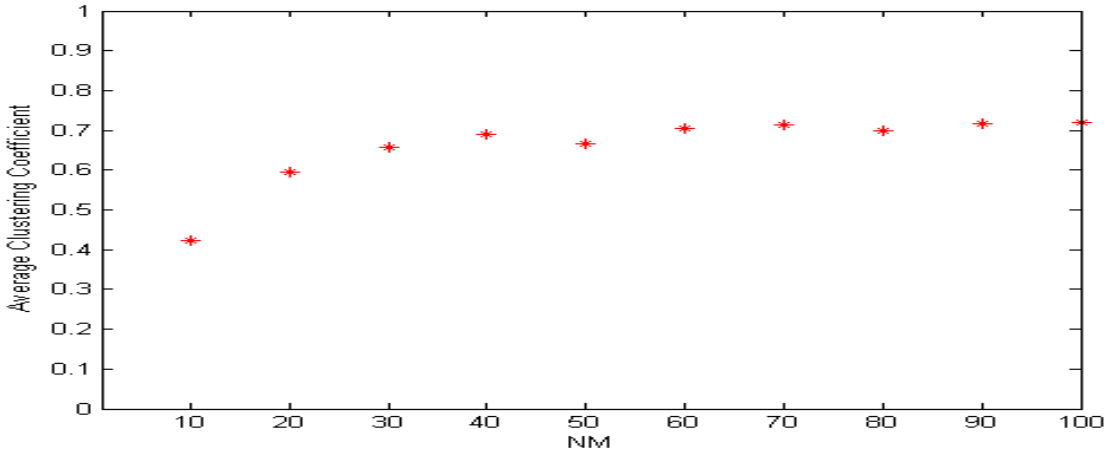
Figure 6.4. SNAM** algorithm with a normalized degree with added attribute similarity with *CF* coefficients $\alpha = 1$, $w = \beta = 0$. a) Average Path Length, b) Power Law Exponent, c) Average Clustering Coefficients



(a) Average Path Length

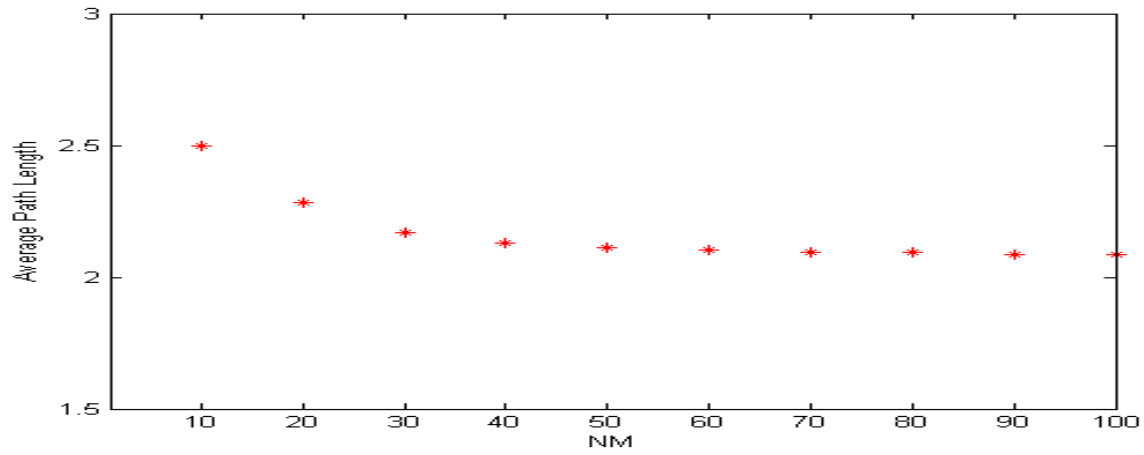


(b) Power Law Exponent

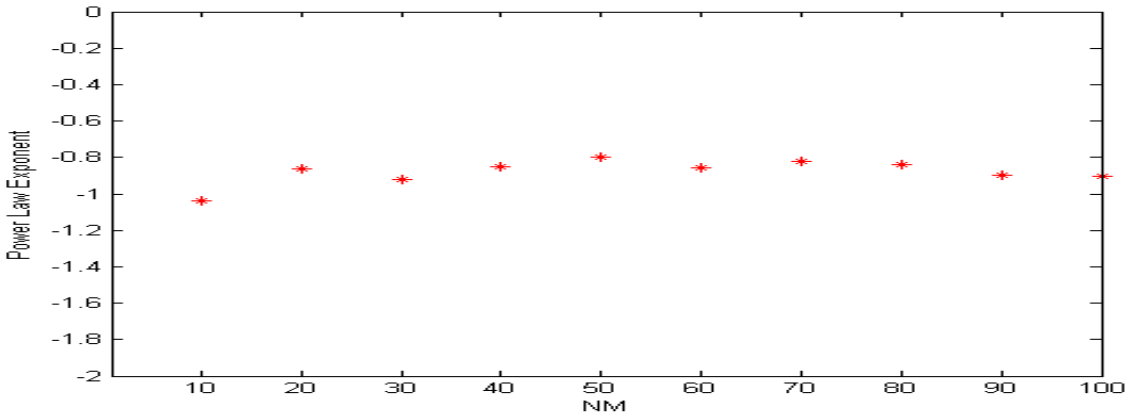


(c) Average Clustering Coefficients

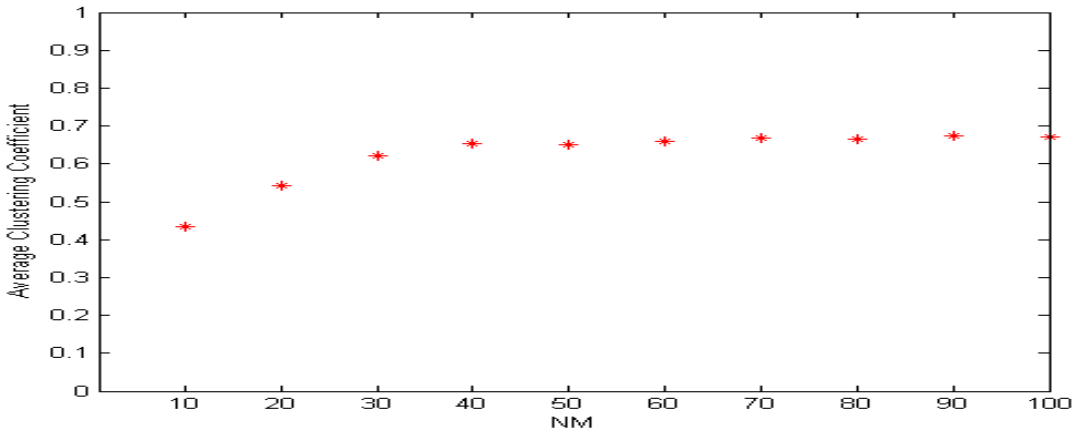
Figure 6.5. SNAM** algorithm with a normalized degree with added attribute similarity with *CF* coefficients $\alpha = 0$, $w = \beta = 0.5$. a) Average Path Length, b) Power Law Exponent, c) Average Clustering Coefficients



(a) Average Path Length



(b) Power Law Exponent



(c) Average Clustering Coefficients

Figure 6.6. SNAM** algorithm with a normalized degree with added attribute similarity with *CF* coefficients $\alpha = 0.8$, $\beta = 0.2$, $w = 0.$:a) Average Path Length, b) Power Law Exponent, c) Average Clustering Coefficients

Table 6.2. Simulation Results for SNAM* for Different Values of *R*

R	**5**	**10**	**15**	**20**	**24**	**25**	**30**	**35**	**40**	**50**	**100**
Exp_PL	-1.6648	-1.6529	-1.8628	-1.762	-1.7578	-1.7545	1.7581	1.7611	-1.7612	-1.7598	-1.7597
Av_Pl	2.4081	2.4110	2.3761	2.0740	2.0677	2.0655	2.0757	2.0726	2.0818	2.0792	2.0745
Av_CC	0.0275	0.0301	0.0585	0.5591	0.6663	0.6712	0.6347	0.6338	0.6328	0.6608	0.6542

Table 6.3. Simulation Results for SNAM** for Different Values of *R*

R	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**50**	**100**
Exp_PL	-1.2961	-1.3248	-1.0964	-1.6005	-2.0808	-2.1363	-2.1362	-2.1381	-2.1320	-2.1396
Av_Pl	4.6842	4.2524	3.1216	2.1242	2.0056	2.0041	2.0026	2.0016	2.0010	2.0024
Av_CC	0.0850	0.1587	0.3610	0.6682	0.6822	0.6791	0.6792	0.6881	0.6904	0.6797

The results in Tables 6.2 and 6.3 indicate that for reduction limit *R* close to or higher than 20, although the value of average clustering coefficient is high, the values of the average path length and the power law exponent are not in the ranges of those of the Facebook dataset used for the case study. On the other hand, values of the statistical averages for a value of *R* around 15 are all in the range of those of the Facebook dataset for this combination of the model's parameters. Changing model parameter values, such as *m*, *NM*, and *CF* coefficients, would generate networks with different statistical properties values for the same *R* value.

6.4.4.2. Analysis of Simulation Results

For SNAM*, the value of the reduction limit *R* affects the network growth mechanism and, thus, it affects values of the statistical parameters of the generated networks. Having a low value for *R* increases the number of random connections that the new nodes have to make to complete their *m* connections after exhausting the *R* reductions. Thus, having a small *R* value can hinder SNAM from performing its connection mechanism that depends on the new node looking for existing nodes that are the most similar to it and then settling for less similar existing nodes. With small *R* values, the new node settles with connections with random nodes after only a few tests which affects the statistical properties of the generated network.

However, a larger value for *R* gives the new node a better chance of finding existing nodes that are more similar to the new node. Results show that a reduction limit less than 20, given the values of other SNAM model parameters used in the simulation, generates a network with low average clustering coefficient values. Additionally, results show that for values of *R* above a certain threshold (in the present set of results the threshold value is $R = 24$), an increase in *R* has only minor effects on the statistical properties of the generated network. Moreover, increasing *R* results in an increase in the run time of the MATLAB program used for network generation.

For SNAM**, results indicate that number of nodes that have to make only one additional random connection is large for small values of *R*. Thus, a large portion of nodes end up having low degree values. As *R* increases, the number of these single random connection nodes decreases since nodes with an excessive reduction of their standard are able to complete their required *m* connections.

6.5. Conclusions of the Case Study

The case study described in this chapter attempted to validate the IASM and SNAM models through a real-world network case study. For starting such validation one must first determine the potentially unique statistical properties of the original network. A small sample of the dataset network was taken as the seed for growing the synthetic network.

Table 6.4 shows the values of the statistical properties power law exponent, average path of the original Facebook dataset and that for networks generated by IASM_A, SNAM* and SNAM** for *CF* with coefficients ($\alpha = 1$, $w = \beta = 0$). The results shown for SNAM* and SNAM** are for

NM = 10 and *R* = 24 and 16. IASM was found incapable of growing networks having statistical parameters values in the same ranges as the real dataset as shown from the subset of results in Table 6.4. IASM_A with and without TFS generates networks with power law exponents of magnitude higher than that of the dataset. Also, IASM_A with and without TFS generates networks with shorter path lengths than that of the dataset. IASM_A without TFS produces lower average clustering coefficients than that of the dataset. The addition of TFS improved only the average clustering coefficient, but it was still less than that of the Facebook dataset.

Results for SNAM were more promising. No results are shown in Table 6.4 for the original SNAM as the algorithm ran indefinitely because of some nodes having null or low attribute similarity values with almost all preexisting nodes.

Failure of our first trials to generate the network using the original SNAM forced a deeper study of the attribute vectors of the nodes in the dataset. The attribute vectors had distributions different from the uniform distribution of the original SNAM model. Investigating the degree distribution of the nodes in the dataset indicated the presence of low degree nodes. Therefore, nodes are allowed to have fewer than *m* connections upon birth as part of extended modified SNAM model (SNAM**). SNAM** generated networks with power law exponent, average path lengths and average clustering coefficients values closer than that of SNAM* to that of the original dataset.

Table 6.4. Statistical properties for original real dataset versus properties of networks generated by IASM_A, SNAM*and SNAM** for *CF* with coefficients ($\alpha = 1$, $w = \beta = 0$).

	Exp_PL	**Av_Pl**	**Av_CC**
Dataset properties	-1.1697	3.6925	0.6055
IASM_A	-1.7620	2.6183	0.0169
IASM_A with TFS	-2.2040	2.7527	0.2213
SNAM* with *NM* = 10, *R* = 24	-1.69	2.09	0.46
SNAM with *NM* = 10, *R* = 16**	-0.8	2.72	0.39

Motivated by our present case study, it is advisable that one starts with thorough study of attribute vectors as well as the node degree distribution of the real network to be modeled. Then the average node degree can be calculated to be considered as parameter *m* of the IASM and

SNAM models. A sample of the real network having size greater than or equal to *m* is used as a seed network to start network growth. The model is used to grow a network to the same size as the dataset. It is advisable to compare growth and dataset network degree distributions and statistical averages.

If the original SNAM does not give statistical averages as those of target network, a deeper study of the adjacency matrix and node attribute vectors should be undertaken to determine possible modifications in the algorithm.

The case study was beneficial in validating the potential use of SNAM in modeling online social networks. The case study shows some guidelines that can be used for modeling other networks. Looking for unique values in the adjacency matrix and the attribute vector for the dataset help us to approach characteristics of the network to be modeled. A reduction limit *R* can be used in the connection algorithm if some nodes have unique attribute values. The degree distribution of the dataset should be studied to see if some nodes possess degree values less than *m*. The choice of *R* depends on the statistical properties desired and algorithm running time. As in the original SNAM, the higher the *NM* the higher the average clustering coefficient and the shorter the average path length of the generated network.

SNAM, being a multi-variable model, gives the opportunity to generate multiple networks in a relatively short time and to find the one that is closest to the original dataset (or that deviates from the original in some desired way). The SNAM connection algorithm is flexible in that it allows slight variations that can reflect other characteristics that are not found in the original model.

Chapter 7. Conclusions and Future Work

This chapter presents the conclusions from the work presented in this dissertation and suggests possible future work.

7.1. Conclusions

This research is based on the premise that mathematical models used to generate complex networks should replicate the statistical properties of real-world networks and should reflect the heterogeneous nature of nodes in real-world networks. We proposed several mathematical models that pave a path to finding a model that successfully reflects the statistical properties of real-world complex networks exhibiting assortative mixing.

The proposed models have heterogeneous network nodes with distinct attributes assigned to different nodes. We modified the network graph $\{V, E\}$ to be of the form $\{V, E, A\}$, where the additional vector A is the attribute set of each network node. To our knowledge, our work is the first to assign more than one attribute to each node. The models are also general in that they do not make any assumptions about the particular type of network modeled.

In Chapters 3 and 4, we first introduced IASM which, to our knowledge, is the first model for the generation of complex networks that integrates a measure for the similarity of attributes with a measure for structural popularity within the connection function, *CF*. In the IASM_A model, nodes are linked preferentially based on a *CF* that depends on the degree of the existing node together with the similarity of attributes between the new node and the existing node. The IASM_B model replaces the measure of node's structural popularity (node degree) used in the IASM_A model with another measure of a node's structural popularity which is based on eigenvector centrality. Both models reflect some statistical properties of real-world complex networks. IASM preserves the power law degree distribution and the small world phenomenon, but the model does not reflect the high average clustering coefficient and the emergence of community structure. We enhanced the IASM model by adding a triad formation step which results in increasing the value of clustering coefficients. Our work on the theoretical algorithm for IASM and its simulation results were published in [46].

Additionally, in Chapters 3 and 4, we introduced another concept of node heterogeneity, which is the heterogeneity of the nodes' requirements to establish a connection. Accordingly, we

proposed a new model, SNAM, which, to our knowledge, is the first model where newly arriving nodes have different connection standard requirements. SNAM is promising as it generates a network that has a power law degree distribution with exponents having magnitudes similar to those found in real-world networks, small average path length, and high clustering coefficient values. SNAM is a general model and excels in its capability to generate various types of complex networks by varying the model parameters. Our work on the connection algorithm for SNAM and the effect of the varying the values of the model parameters were published in [47] and [48].

In Chapter 4, we investigated the presence of community structure in both of our models, IASM and SNAM. We modified the connection algorithm of IASM and SNAM to induce the emergence of community structure by adding a class similarity coefficient, μ, in the connection function, *CF*. The community structure was examined by finding the percentage of inter-class connections. Networks generated using these models show community structure as most connections are made between members of the same class. The proposed community structure models preserve the power law degree distribution and the small world property of complex networks. There was also a slight decrease in the average clustering coefficient.

We next presented an analytical representation for the degree distribution for networks generated by IASM and SNAM in Chapter 5. We used the rate equation for the degree of the new node arriving to the network. Then this rate equation was processed mathematically until a mathematical expression for the degree distribution was found for special cases of IASM and SNAM. IASM_A with *CF* depending on degree centrality multiplied by attribute similarity and SNAM with *CF* depending only on the degree centrality were proven analytically to generate networks with a power law degree distribution. Our additional results for the SNAM community structure model and the mathematical analysis of SNAM are expected to appear in a book chapter [49].

Our proposed models are general and, hence, can be used to generate any type of complex network. As a proof of concept, we considered a case study in Chapter 6where these models are applied to online social networks. Our choice of online social networks is mainly due to their widespread use and current interest in applications in many fields such as marketing, information diffusion, recommendation, and trust analysis. We use a dataset based on social networks from

Facebook to validate the potential of our models in generating networks that mimic such a real-world complex network. It was found that the characteristics of the network generated by IASM were different from that of the original dataset. The original SNAM was inadequate in modeling this dataset because of the existence of nodes with unique attribute values. Additionally, upon examining the degree distribution of the original dataset, it was found that some nodes have degree centrality values less than *m*, where *m* is the number of links each arriving node has to establish in SNAM. Thus, two modified versions of SNAM, SNAM* and SNAM**, were introduced. Extended SNAM** has a reduction limit parameter, *R*, that limits the number of reductions that a new node makes to find nodes satisfying its connection standard. Additionally, upon reaching its reduction limit *R*, the new node only makes one additional connection to a random node rather than completing its *m* connections as in the case of SNAM*. Extended SNAM** was successful in generating networks having statistical properties values near that found in the original Facebook dataset. The case study has provided us with some useful guidelines for adapting the proposed SNAM model to more accurately model real-world complex networks with different characteristics.

In our motivating goals, we stated that our models should preserve the four statistical properties common to real-world complex networks. We were able to validate that the four statistical properties exist in the networks generated by IASM and SNAM using simulation, mathematical analysis, and/or a case study. The methods used to demonstrate the properties are presented in Table 7.1. If the statistical property was validated for a special case of the model, an asterisk (*) is added beside the name of the validation method in Table 7.1. As seen in Table 7.1, the power law degree distribution is validated for both models using all three methods. The existence of high average clustering coefficients was validated for both SNAM and IASM with a TFS via simulation and the case study. Additionally, the small world phenomenon or having small average path lengths was validated using simulation and the case study for both IASM and SNAM. Finally, the emergence of the community structure was validated only via simulation.

Additionally, the nodes presented in our models were heterogeneous. This is validated in the case study by the ability of IASM and SNAM of incorporating the attribute vectors of the dataset's network nodes. There were no assumptions about the nature of the nodes' heterogeneity parameter as was seen also in simulations, the mathematical analysis, and the case study. Thus,

our models are general for any complex networks exhibiting a bias when making connections by the new nodes towards nodes having certain properties (assortative mixing).

Table 7.1. Validation Methods for the Desired Statistical Properties for IASM and SNAM

	SNAM	**IASM**
Power Law Degree Distribution	• Simulation • Mathematical Analysis* • Case study	• Simulation • Mathematical Analysis* • Case study
High Average Clustering coefficients	• Simulation • Case study*	• Simulation (with a TFS) • Case study (with a TFS)
Small Average Path Lengths	• Simulation • Case study*	• Simulation • Case study
Community Structure Emergence	• Simulation	• Simulation

7.2. Future Research Ideas

Our analysis of the BA model revealed unfairness with most of the new added connections being made with older network nodes. This gives the older nodes, which most likely become less active with time, an unfair advantage for connections over newer nodes. The idea of fairly distributing connections among network nodes added at different times across the life of the generated network is of interest as old nodes usually attain most of the new connections. Also, old nodes have a tendency to become less active with time. Thus, the new node gains more advantages by making some of its connections with very old nodes and some with more recent preexisting nodes.

The attributes in our models are as seen abstract. Another potential research direction is to experiment with the nature of attributes. Experimenting with the nature of attributes can be useful in making the generated networks mimic some phenomena that are found in real-world complex networks. We believe that using affinity matrices for attributes in our model will prove fruitful in making the model more lifelike. This reflects the idea that sometimes having opposite attribute-values would encourage users to establish connections. For example, in social networks, a user might be inclined to form connections with users of the opposite gender.

Moreover, the introduction of the time element into our models is also of potential value. The time element can be introduced by adding the time element as one of anode's attributes. This addition could be useful in generating time sensitive networks. Time sensitive networks include networks where the arrival time of the node can affect its properties and its future connections. Time sensitive attributes can be used to represent nodes becoming less active with time or nodes changing their interests and thus their attributes. Another form of time sensitive attributes are networks where nodes change their attributes with time. Investigating the effect of the presence of time sensitive attributes on the generated network structure can be beneficial.

Moreover, the inclusion of new structural phenomena in our models is another research direction. These structural phenomena reflect some behaviors that can be found in real complex networks. First, rewiring based on time sensitive attributes can be used to mimic the situation when nodes terminate connections with old, inactive nodes to form new connections. Second, the assumption made by most models for network generation that each node makes the same number of connections, *m*, is not always true. Each arriving node can have a different *m* value. This value can depend on the structural properties of the network upon the arrival of this node. Thus, varying the number of connections, *m*, that arriving nodes make based on network properties can be investigated.

References

[1] F. Emmert-Streib and M. Dehmer, "Exploring Statistical and Population Aspects of Network Complexity," *PLoS One*, vol. 7, no. 5, e34523, May 8, 2012.

[2] X. F. Wang and G. Chen, "Complex networks: Small-world, scale-free and beyond," *IEEE Circuits and Systems Magazine*, vol.3, no.1, pp. 6-20, September 2003.

[3] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-512, October 1999.

[4] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167-256, March 2003.

[5] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Review of Modern Physics*, vol. 74, no. 1, pp. 47-97, January-March 2002.

[6] H. Balakrishnan and N. Deo, "Discovering Communities in Complex Networks," *Proceedings ACM Southeast Conference*, pp. 280-285, March 10-12, 2006.

[7] E. Ferrara, "Mining and Analysis of Online Social Networks," Ph.D. Dissertation, University of Messina, February 2012.

[8] R. Albert and A.-L. Barabási, "Topology of evolving networks: Local events and universality," *Physics Review Letters*, vol. 85, no. 24, pp. 5234-5237, December 2000.

[9] V. Samalam, "Preferential attachment alone is not sufficient to generate scale free random networks," arXiv:1202.1498 [physics.soc-ph], 4 pages, February 2012.

[10] P. L. Krapivsky, S. Redner, and F. Leyvraz, "Connectivity of growing random networks," *Physics Review Letters*, vol. 85, no. 21, pp. 4629-4632, November 2000.

[11] M.-Y. Wang, G. Yu and D.-R. Yu, "The scale-free model for citation network," *Proceedings IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, vol. 1, pp. 773-776, October 29-31, 2010.

[12] J. Wang, L. Rong, and L. Zhang, "Evolving small-world networks of the local world with tunable clustering," *Proceedings ISECS International Colloquium on Computing, Communication, Control, and Management*, vol. 2, pp. 369-373, August 3-4 2008.

[13] A.-L. Barabási, H. Jeong, E. Ravasz, Z. Néda, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3-4, pp. 590-614, August 2002.

[14] S. N. Dorogovtsev and J. F. F. Mendes, "Scaling behaviour of developing and decaying networks," *Europhysics Letters*, vol. 52, no. 1, pp. 33-39, August 2000.

[15] J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The Web as a graph: Measurements, models and methods," *Proceedings International Conference on Combinatorics and Computing*, *Lecture Notes in Computer Science*, no. 1627, pp. 1–18, Springer, Berlin, 1999.

[16] P. L. Krapivsky and S. Redner, "Organization of growing random networks," *Physics Review E*, vol. 63, no. 6, 066123, June 2001.

[17] C. Herrera and P. J. Zufiria, "Generating scale-free networks with adjustable clustering coefficient via random walks," *Proceedings IEEE Network Science Workshop (NSW)*, pp. 167-172, June 22-24, 2011.

[18] D. Wang, X. Qian, and X. Jin, "Dynamical evolution of weighted scale-free network models," *Proceedings Chinese Control and Decision Conference (CCDC)*, pp. 479-482, May 23-25, 2012.

[19] L.-R. Wu and Q. Yan, "Modeling dynamic evolution of online friendship network," *Communications in Theoretical Physics*, vol.58, no. 4, pp. 599-603, October 2012.

[20] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical Review E*, vol. 65, no. 2, p. 814-822, February 2002.

[21] S. Bhukya, "A novel model for social networks," *Proceedings Baltic Congress on Future Internet Communications*, pp.21-24, February 16-18, 2011.

[22] P. Fu and K. Liao, "An evolving scale-free network with large clustering coefficient," *Proceedings International Conference on Control, Automation, Robotics and Vision*, pp. 1-4, December 5-8, 2006.

[23] J. Wang and L. Rong, "Evolving small-world networks based on the modified BA model" *Proceedings International Conference on Computer Science and Information Technology*, pp. 143-146, August 29-September 2, 2008.

[24] J.-G. Liu, Y.-Z. Dang, and Z.-T. Wang, "Multistage random growing small-world networks with power-law degree distribution," *Chinese Physics Letters*, vol. 23, no. 3, p. 746-749, October 2006.

[25] K. Klemm and V. M. Eguiluz, "Growing scale-free networks with small world behavior," *Physical Review E*, vol. 65, 057102, May 2002.

[26] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings National Academy of Sciences of the United States of America*, vol. 99 (Supplement 1), pp. 2566-2572, February 2002.

[27] S. N. Dorogovtsev and J. F. F. Mendes, "Growing network with heritable connectivity of nodes,"arXiv:cond-mat/0011077 [cond-mat.stat-mech], 4 pages, November 2000.

[28] H. Li, H. Zhao, W. Cai, J.-Q. Xu, J. Ai, "A modular attachment mechanism for software network evolution," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 9, pp. 2025-2037, May 2013.

[29] F. Zaidi, A. Sallaberry, and G. Melancon, "Generating artificial social networks with small world and scale free properties," INRIA Research Report RR-7861, 2012, 34 pages, https://hal.inria.fr/hal-00659971, Accessed May 1, 2015.

[30] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Physics Review Letters*, vol. 85, no. 20-21, pp. 4633-4636, November 2000.

[31] S. N. Dorogovtsev and J. F. F. Mendes, "Effect of the accelerating growth of communications networks on their structure," *Physical Review E*, vol. 63, no. 2, 025101, February 2001.

[32] G. Bianconi and A.-L. Barabási, "Competition and multiscaling in evolving networks," *Europhysics Letters*, vol. 54, no. 4, pp. 436-442, May 2001.

[33] R. Sun, W. Luo, A. Mu, L. LI, and M. Zhong, "Complex network model based on node attraction with tunable parameters," *Proceedings International Conference on Network and Computational Intelligence*, pp. 11-16, August 3-4, 2012.

[34] G. Cai, R. Wang, and B. Qiang, "Online social network evolving model based on damping factor," *Procedia Computer Science*, vol. 9, pp. 1338-1344, 2012.

[35] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the National Academy of Science of the United States of America*, vol. 97, no. 21, pp. 11149–11152, October 2000.

[36] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of reference networks with aging," *Physical Review E*, vol. 62, no. 2, p. 1842-1845, August 2000.

[37] X. Zhang, T. Chen, R. Chen, and H. Li, "Complex network modeling with constant capacity restriction based on BA Model," *Proceedings International Symposium on Computer Network and Multimedia Technology*, pp. 1-4, January 18-20, 2009.

[38] S. Tao and X. Yue, "The attributes similar-degree of complex networks," *Proceedings International Conference on Future Computer and Communication*, vol. 3, pp. 531-535, May 21-24, 2010.

[39] M. Kim and J. Leskovec, "Multiplicative attribute graph model of real-world networks," *Proceedings Seventh Workshop on Algorithms and Models for the Web Graph*, 2010.

[40] Y. Li, X. Jin, F. Kong, and J. Li, "Linking via social similarity: The emergence of community structure in scale-free network," *Proceedings IEEE Symposium on Web Society*, pp. 124-128, August 23-24, 2009.

[41] MATLAB, http://www.mathworks.com/products/matlab/, Accessed May 1, 2015.

[42] A.-L. Barabási, R. Albert, H. Jeong, "Mean-field theory for scale-free random networks," arXiv:cond-mat/9907068 [cond-mat.dis-nn], 19 pages, July 1999.

[43] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol.13, no. 1, pp. 210-230, October 2007.

[44] L.A. Cutillo, R. Molva, and T. Strufe, "Privacy preserving social networking through decentralization," *Proceedings International Conference on Wireless On-Demand Network Systems and Services*, pp. 145-152, February 2-4, 2009.

[45] SNAP–Social Circles: Facebook, http://snap.stanford.edu/data/egonets-Facebook.html, Accessed May 1, 2015.

[46] B. E. Youssef and H. Hassan, "IASM: An integrated attribute similarity for complex networks generation," *Proceedings IEEE International Conference on Information Networking (ICOIN)*, pp. 567-571, February 10-12, 2014.

[47] B. E. Youssef and M. R. M. Rizk, "Effect of arriving nodes connection-standards on models for the generation of heterogeneous complex networks" *Proceedings Workshop on Complex Systems Modeling and Simulation (CoSMoS'14)*, pp. 13-34, July 30, 2014.

[48] B. E. Youssef and M. R. M. Rizk, "SNAM: A heterogeneous complex networks generation model," *Proceedings International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine)*, pp.44-50, August 18-20 2014.

[49] B. E. Youssef, "SNAM: A Heterogeneous Complex Networks Generation Model," chapter in *Advanced Methods for Complex Network Analysis*, submitted.