CENTRALITY MEASURES ON NETWORKS AND EMPIRICAL ANALYSIS ON ACTIVITY DRIVEN NETWORK MODELS

by

ECE NAZ DUMAN

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of

the requirements for the degree of

Master of Science

Sabancı University

August 2015

CENTRALITY MEASURES ON NETWORKS AND EMPIRICAL ANALYSIS ON ACTIVITY DRIVEN NETWORK MODELS

APPROVED BY:

Prof.Dr.Ali Rana Atılgan.....
(Thesis Supervisor)

Assoc.Prof.Dr. Güvenç Şahin.....

Asst.Prof.Dr.Ahmet Onur Durahim.....

DATE OF APPROVAL: 05.08.2015

ABSTRACT

CENTRALITY MEASURES ON NETWORKS AND EMPIRICAL ANALYSIS ON ACTIVITY DRIVEN NETWORK MODELS

ECE NAZ DUMAN

M.S. in Industrial Engineering

Master Thesis, August 2015

Thesis Supervisor : Prof.Dr. Ali Rana Atılgan

Keywords: Power Laws, Evolving Network, Triadic Closures, Centrality Indicators

Social network analysis involves structural studies on social networks, and it benefits measures of graph theory. Centrality indicators are one of these measures, and they map the characteristics of networks. In the first part of this thesis we study the effects of centrality measures on various network types, including degree, closeness and betweenness centrality.

Connections in social networks are rapidly changing, through triadic closures, membership closures or foci closures. Correspondingly, the second part of this thesis and the main objective is first to imitate modeling of an evolving network which thoroughly repeats the behavior of real life networks, and then study triadic closures and topological features of this model. Generation of this model includes power law distribution; thus, its degree distribution and other properties resemble features of three studied datasets. Our analysis on this model gives novel results since the model forms triadic closures as in actual social networks. At the end of our analysis we discuss the reasons of triadic closures in this model with the help of centrality indicators and clustering coefficient.

ÖZET

AĞLARDA MERKEZİYET ÖLÇÜLERİ VE AKTİVİTEYE DAYALI AĞLARDA DENEYSEL ANALİZLER

ECE NAZ DUMAN

Endistri Mühendisliği Yüksek Lisansı

Yüksek Lisans Tezi, Ağustos 2015

Tez Danışmanı: Prof.Dr. Ali Rana Atılgan

Anahtar Kelimeler: Kuvvet Yasası, Gelişen Ağlar, Üçlü Kapanımlar, Merkeziyet Ölüleri

Sosyal ağ analizleri, sosyal ağlarda yapısal çalışmaları içerir ve grafik teorisi ölçülerinden faydalanır. Merkeziyet ölçüleri bu ölçülerden bir tanesidir ve onlar ağların karakteristik özelliklerini göstermede kullanılırlar.Bu tezin ilk kısmında derece dağılımı, yakınlık ve arada olma gibi merkeziyet ölçülerinin farklı ağ tiplerine olan etkilerini inceleyeceğiz.

Sosyal ağlardaki bağlar üçlü kapanımlar, üyelik kapanımları veya ortak odak noktaları aracılığıyla sürekli olarak değişmektedir. Buna bağlı olarak, bu tezin ikinci kısmı ve ana amacı öncelikle gerçek hayat ağlarının davranışlarını oldukça iyi taklit eden bir gelişen ağın modellemesini tekrar etmek ve daha sonra bu ağın topolojik özelliklerini ve üçlü kapanımlarını analiz etmektir. Bu ağın modellenmesi kuvvet yasası dağılımını kapsamaktadır. Dolayısıyla ağın derece dağılımı ve diğer özellikleri daha önce araştırılmış olan üç data setinin özelliklerine benzemektedir. Bu ağ üzerindeki analizlerimiz eşi benzeri olmaya sonuçlar vermektedir çünkü bu ağ, hakiki sosyal ağlardaki gibi üçlü kapanımlar oluşturmaktadır. Analizlerimizin son kısmında merkeziyet ölçülerinin ve kümelenme katsayısının yardımıyla bu ağdaki üçlü kapanımların sebeplerini tartışacağız.

This humble work is dedicated to

my beloved sister Dila

&

my precious Sarp

&

My dearest friends

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my thesis supervisor Ali Rana Atılgan for providing me his advice and constant encouragement throughout the course of this thesis.

I would also like to thank Güvenç Şahin and Ahmet Onur Durahim for accepting to be part of thesis jury and their valueble feedback.

I gratefully acknowledge the funding received from TÜBİITAK BİDEB to complete my master degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social network analysis is an interdisciplinary area of research, which is bursted from network science, graph theory and social sciences. In social networks a node represents a social actor such as a person, an affiliation, an edge shows the communication between two people, a membership connection of a person to a foci and flow represents a message, a phone call and etc. Furthermore, usually social networks are undirected graphs.

Understanding a social network's structure has various applications in areas ranging from marketing to disease transmission. Thus, social network analysis benefits methods of graph theory for studying structure of social networks. Additionally, it provides many measures to examine a nodes position in a network, how information flows through it, and its local structure.

One of the measures, which are utilized for studying a network's structure, is centrality, a set of indicators that determines the importance of vertices in a network. Researchers have been using centrality to understand the characteristics of a network, and they are applied for solving common problems of social networks [28]. There are many centrality indicators, and each one defines a vertex's centrality, but the most commons are degree, closeness and betweenness centrality. The first one, degree centrality, is the simplest of all, and it specifies number of

connections each node has in a network [1]. The second one, closeness centrality of a vertex, also known as average path length, is its mean shortest-path distance from all other vertices [22]. The last one, betweenness centrality is also based on network's paths. However, a node's betweenness centrality is not about its distance from other nodes but about the fraction of paths that it lies on. If the information flows through geodesic paths, *shortest path betweenness* is preferred for analyses [11]. On the other hand, if the given information flows along the network randomly like a rumor, without any target, then *random walk betweenness* can be used to study centrality of the vertices of a network [20].

In some cases inquiring centrality of a vertex in a network is inadequate for exploring its global structure. Knowing connectivity pattern of a network provides many advantages for estimating how the information flows on it [8]. One way to analyze the connectivity pattern of a network is benefiting from clustering coefficient [13]. Clustering coefficient of a vertex determines the fraction of its friends that are connected to each other.

Furthermore, network scientists have been interested in topic network modeling [26] [19]. Well-known network models like scale-free networks and random graphs are studied many times [3] [9]. Additionally, evolving networks attract great attention recently; since it is observed many times in social networks that vertices or links can be added to or removed from the network over time [14]. One algorithm for generating a growing network is presented by Perra et al. [24], and the model is called activity driven network. This algorithm offers a network which is significantly similar to real life social networks. Thus, it can be implemented for applications like contagion diseases [25] and marketing problems without the need of data analysis.

Social networks are evolving networks since individuals either build new connections in time or they lose contact with some old friends. An investigation on evolving social networks is about the factors which are effective on forming a new link in a given social network [17] [23]. These factors can be personal factors, such as changing job, neighborhood or interests, or more general factors such as having mutual friends or common affiliations. If two individuals connect with an edge since they have at least one friend in common, this formation is called triadic closure. It is proven for several datasets that the probability of triadic closures is affected from the number of common friends of two people [17] [23].

The main objectives of this thesis are to observe the effect of number of common friends on link formation for a network model that is generated as Perra et. al. [24] suggested and to analyze the reasons behind its results.

Rest of the paper follows the given structure: First we will introduce centrality measures in chapter 2 in the interest of understanding why they are needed, observing their applications on well-known networks, and analyzing correlations among them. Next, we will explain generation of activity driven networks in detail, and observe their centrality indicators and clustering coefficient. Last, in chapter 5 we will explore triadic closures in activity driven networks in case that two individuals have multiple number of friends in common. Moreover, we will favor results of chapter 4 to explain outcome of chapter 5.

Chapter 2

Betweenness Centrality Measures and Their Correlations

In social network analysis calculating importance of vertices helps researchers to understand the fundamentals of a network. Most common applied tools which indicate importance of a vertex are centrality indicators. Researchers of social network science offer many indicators that define a vertex's centrality in a network. For example, a vertex with a relatively high degree centrality has more neighbors in compared to others. The one with the highest closeness centrality has smallest average distance to all other nodes in terms of number of links. On the other hand betweenness centrality of a vertex defines its centrality with respect to which paths it stands on. In this chapter we will investigate two betweenness centrality indicators for different network types.

2.1. Shortest-Path and Random-Walk Betweenness for Social Networks

2.1.1. Shortest-Path Betweenness

We will start with the earlier and most common version of betweenness centrality is defined by Freeman (1977) [11], and we will refer to this as shortest path betweenness. Considering two vertices in a network vertex s and vertex t, shortest path betweenness of a vertex i is fraction of shortest paths from vertex s to vertex t for all s and t, which passes through vertex i. Since we work with homogeneous networks in our research, on which strength of the links are binary, a shortest-path means a path with minimum number of links. Specifically, let $g_i^{(st)}$ be the number of shortest paths from source vertex s to target vertex t that passes through vertex i, number of paths that includes minimum number of links, and let η_{st} be total number of shortest paths from s to t. Then one can use the equation (1) below to calculate the betweenness of vertex i [11].

$$b_i = \frac{\sum_{s<t} g_i^{(st)} / \eta_{st}}{\frac{1}{2} n(n-1)} \tag{1}$$

In order to demonstrate an example we measured the shortest path betweenness of well-known networks like scale-free networks and Erdős–Rény random graphs. In figure 1 we analyze the average of shortest path distributions of 20 random graphs with probability of existing in the network for each possible edge p=0.4, and number of nodes N=500.

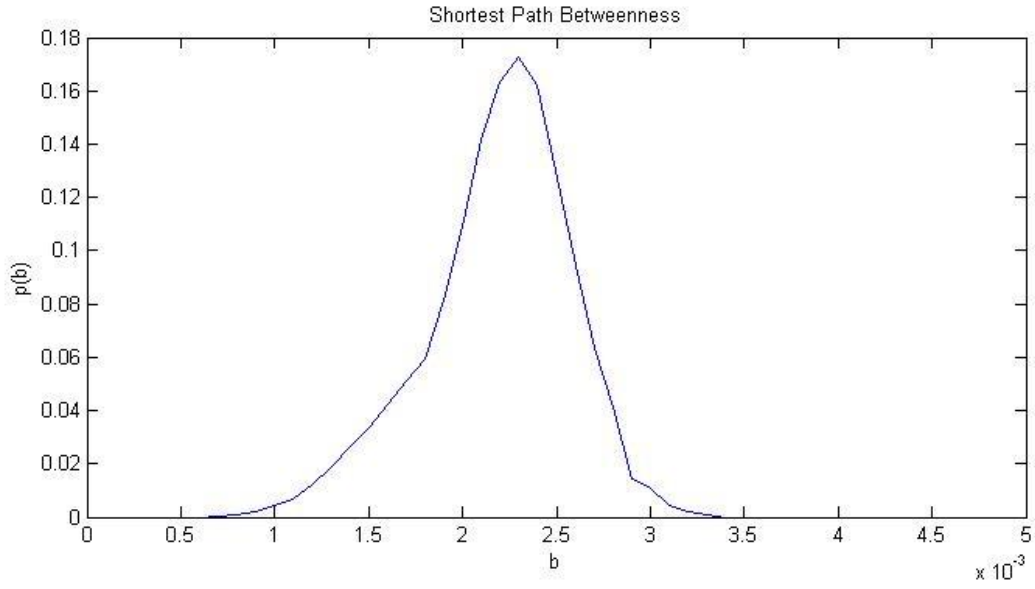


Figure 1: Average Shortest Path Betweenness Distribution of 20 Random Graphs

Shortest path betweenness distribution of random graphs is symmetric. It has a passion shape shows that shortest path betweenness of most nodes is around the mean value, and few nodes have very low or very high betweenness values. Moreover, it is not possible to observe extreme values of betweenness.

Additionally, average shortest path betweenness distribution of 20 scale free network with 600 nodes, is given in figure 2.

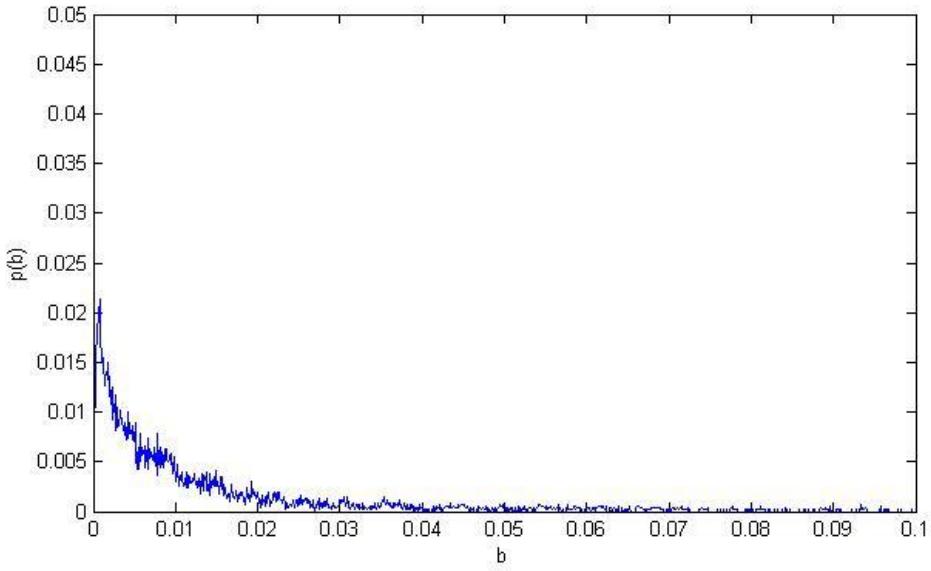


Figure 2: Average Shortest Path Betweenness Distribution of 20 Scale-free Networks

This plot has a skewed shape and it seems like the shape of power law distribution. There are a few nodes with extreme high betweenness and many nodes with extreme low betweenness. From chapter 3 (see page 21) we know that power laws have a linearly decreasing shape when we take the logarithm of both axes. Hence to check whether scale free networks have a power- law distributed shortest path betweenness or not we also examined this plot in logarithmic scale in figure 3.

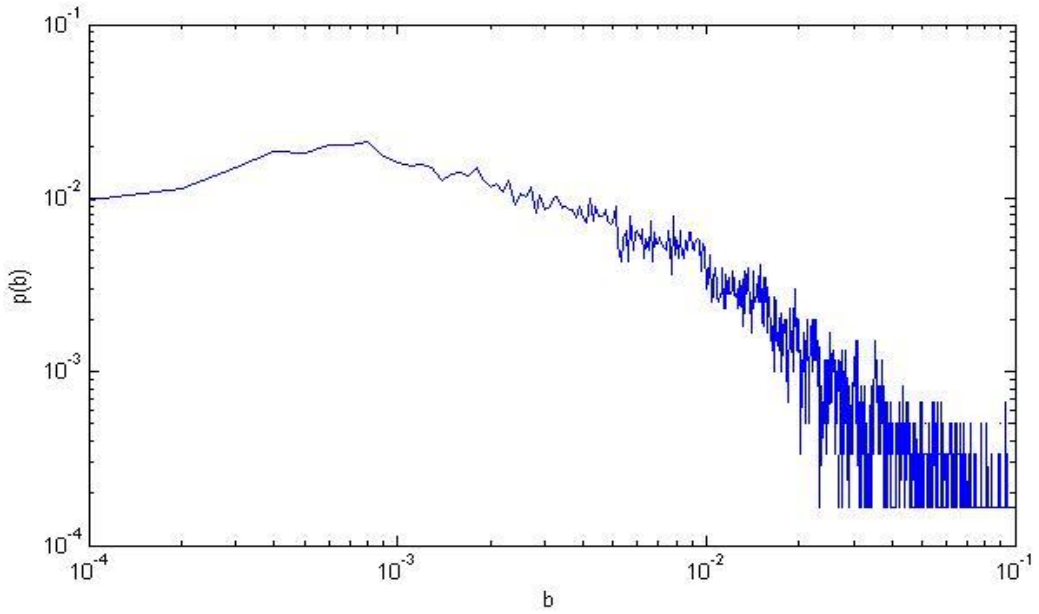


Figure 3 : Average Shortest Path Betweenness Distribution of Scale-free Networks in Logarithmic Scale

The shape of the plot is not linearly decreasing in all areas, for small values of b until $b \approx 10^{-3}$ the slope has a positive value and the curve moves upwards. Then, the probability of observing higher values of shortest path betweenness decreases linearly up to 0.02 and after this point the curve looks still power-law but gets noisier since number of nodes with such high betweenness values declines. Comparison of figure 1 and 3 demonstrates us while the majority of shortest path betweenness values of nodes in random graphs are around mean value, for scale free networks most of the nodes have very low shortest path betweenness values.

2.1.2. Random-Walk Betweenness

Unfortunately, shortest path betweenness is inadequate to explain many situations where information like rumor or news does not necessarily flow along the shortest paths. Study of Dodds (2003) [7] states that even if participants are informed to forward a message to a target in the most directed way possible, there is no prove that they will select the shortest paths to send the message. Newman (2005) [20] developed a new method for calculating betweenness centrality based on random walk procedure of a network, which will be referred as random walk betweenness. On homogeneous networks, random walk is a process which can be explained by the movement of an imaginary ant on links of a network, and which wanders around uniform randomly, without any biased on where to go next, after starting at a source node s. Thus, random walk betweenness of a vertex i is the fraction of times that vertex i appears in random walks between all pairs of source and target nodes, averaged over for many number of random walk trials.

Clearly, if the ant crosses a vertex back and forth during random walk, its betweenness value will be high although he does not change his location. Therefore we assume passing through a vertex and then passing back through it cancels out in random walk betweenness calculations.

Complexity of random walk betweenness in a network is O $((m+n)*n^2)$ with matrix methods established by Newman and Girvan (2004) [21]. We also followed their steps to calculate random walk betweenness.

In figure 4 and 5 average random walk betweenness distributions of 20 random graphs and scale free networks with the same parameters as above are given.

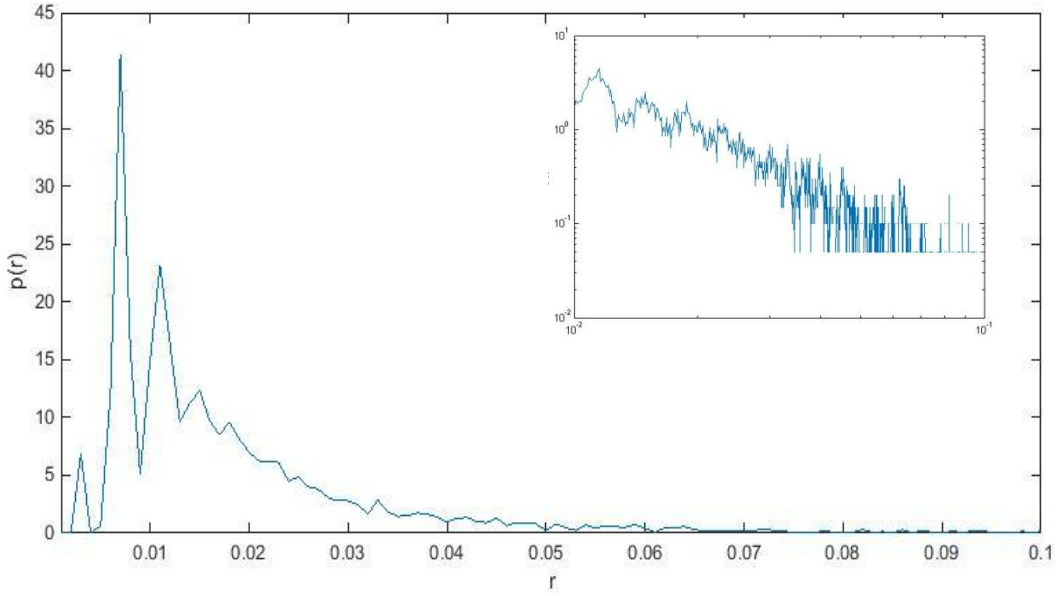


Figure 4: Average Random-walk Betweenness Distribution of 20 Scale-free Networks. Small figure is the Same Plot in Logarithmic Scale

Random-walk betweenness curve of scale-free networks in figure 4 appears to have a slight different shape than shortest-path betweenness of them in figure 3, whereas shape of two plots in figures 1 and 5 are more similar to each other. In the next section we will quantify this similarity.

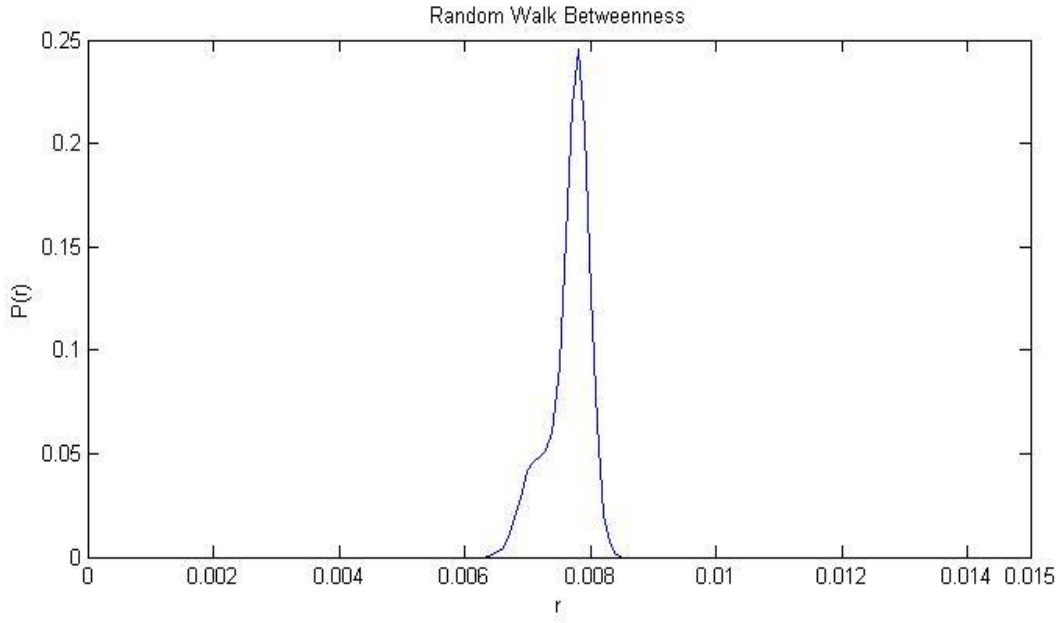


Figure 5: Average Random-walk Betweenness Distribution of 20 Random Graphs

2.2. Correlations among Betweenness Measures

After the analogy of shapes of random-walk betweenness and shortest-path betweenness curves we will examine another research area about betweenness centrality that involves measuring correlations among different centrality indicators; degree, shortest-path betweenness and random-walk betweenness.

Newman (2005) [20] studied a network presented by Potterat et. al. (2002) [25], and discovered that random walk betweenness is highly correlated with degree (r^2=0.626) and even more correlated with shortest path betweenness (r^2=0.923).

In our project we expand this work and measure the correlations for many other networks. First, we search for correlation in known datasets such as Schwimmer taro exchange network [12] and Kapferer mine network [15]. Next, we include the plots which display correlations for random graphs and scale-free networks with different parameter settings. Schwimmer Taro Exchange dataset represents the relation of gift-giving (taro exchange) among 22 households in a Papuan village. Its random walk betweenness correlation with degree is in figure 6 on the left side, and with shortest path betweenness is in on the right.

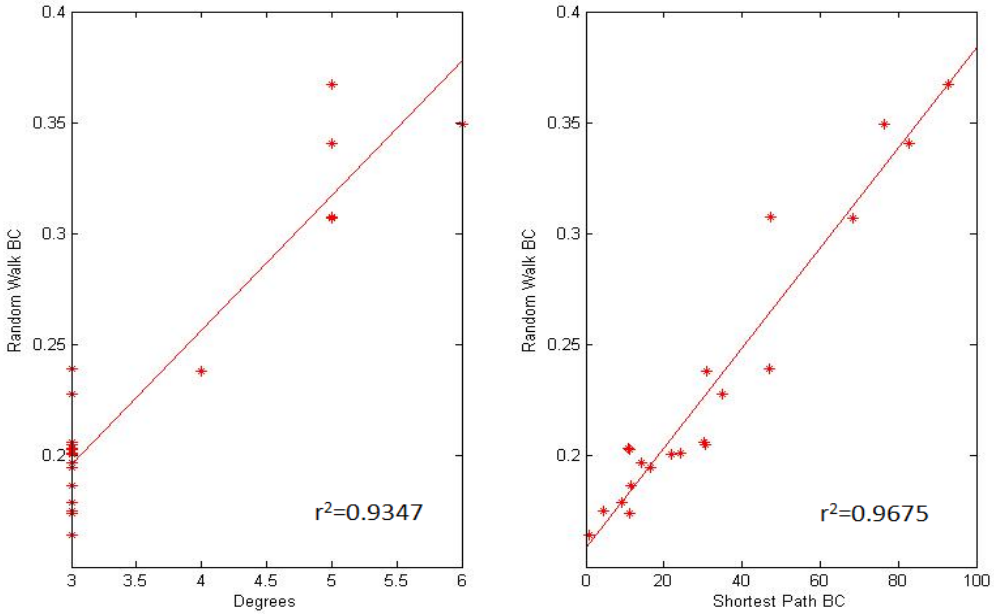


Figure 6: Plots of Correlations RWBD and RWBS for Schwimmer Taro Exchange Dataset

In figure 7 results of Kapferer Mine dataset are given. This dataset represents social interactions among 15 workers of a mining operation in Zambia.

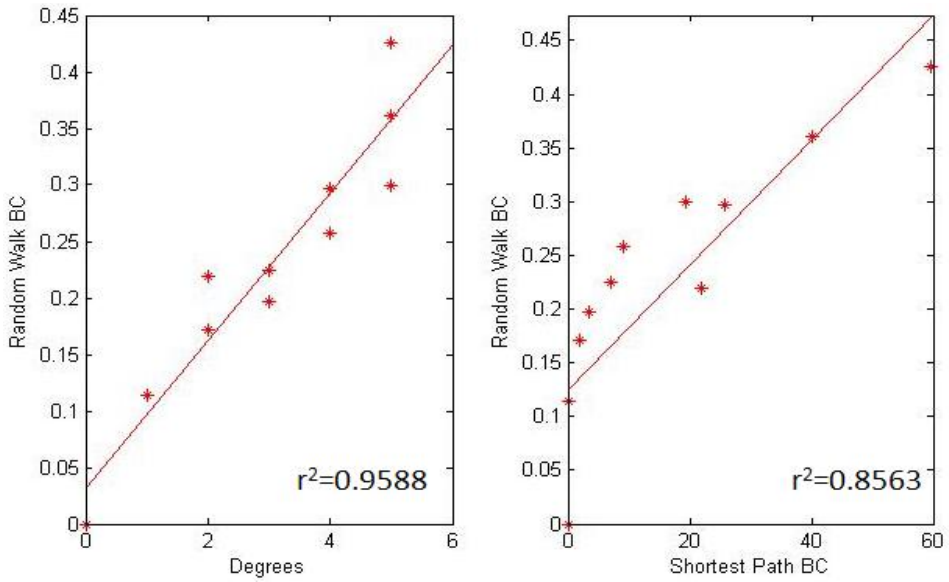


Figure 7: Plots of correlations RWBD and RWBS for Kapferer Mine Dataset

Additionally we observe correlations for scale-free networks and Erdős–Rény random graphs which we generated. Interestingly, for random graphs both correlation scores are very high. Furthermore, random-walk betweenness correlation with degree (we will refer this correlation as RWBD) is higher than correlation with shortest path (for this correlation RWBS will be used) for observed scale-free networks, but the situation is vice versa for random graphs.

First parameter for the generation of G (N, p) model Erdős–Rény random graphs [9] [10] is number of vertices N and the second one is the independent probability of an edge to be included in the graph.

We measured correlation coefficients for several different parameter sets of random graphs, and repeated calculations for five different networks with each parameter set. Parameter sets are given in table 1.

Table 1: Correlation Coefficients of RWBD and RWBS for Random Graphs with Different Settings

N	p	*Network 1*		*Network 2*		*Network 3*		*Network 4*		*Network 5*	
		RWBD	**RWBS**	**RWBD**	**RWBS**	**RWBD**	**RWBS**	**RWBD**	**RWBS**	**RWBD**	**RWBS**
100	0,4	0.9855	0.9918	0.9863	0.9897	0.9900	0.9943	0.9900	0.9937	0.9871	0.9922
200	0,4	0.9941	0.9974	0.9946	0.9973	0.9938	0.9968	0.9952	0.9971	0.9944	0.9971
200	0,5	0.9954	0.9976	0.9946	0.9965	0.9952	0.9975	0.9949	0.9970	0.9957	0.9977
200	0,7	0.9962	0.9973	0.9963	0.9972	0.9963	0.9968	0.9962	0.9970	0.9963	0.9967
300	0,4	0.9960	0.9982	0.9965	0.9982	0.9961	0.9983	0.9961	0.9981	0.9963	0.9985
300	0,5	0.9966	0.9985	0.9970	0.9983	0.9964	0.9986	0.9969	0.9986	0.9969	0.9985
500	0,5	0.9978	0.9990	0.9980	0.9992	0.9978	0.9992	0.9979	0.9991	0.9978	0.9992

We draw in figures 8, 9 and 10 one plot for each parameter setting given in tableau 1.

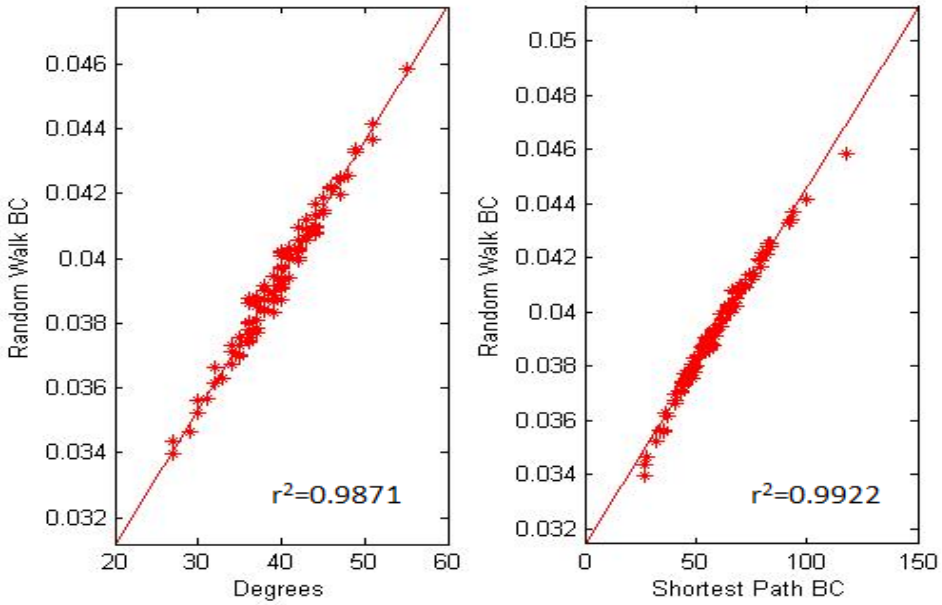


Figure 8 : Plots of Correlations RWBD and RWBS for a Random Graph with p=0.4 and N=100

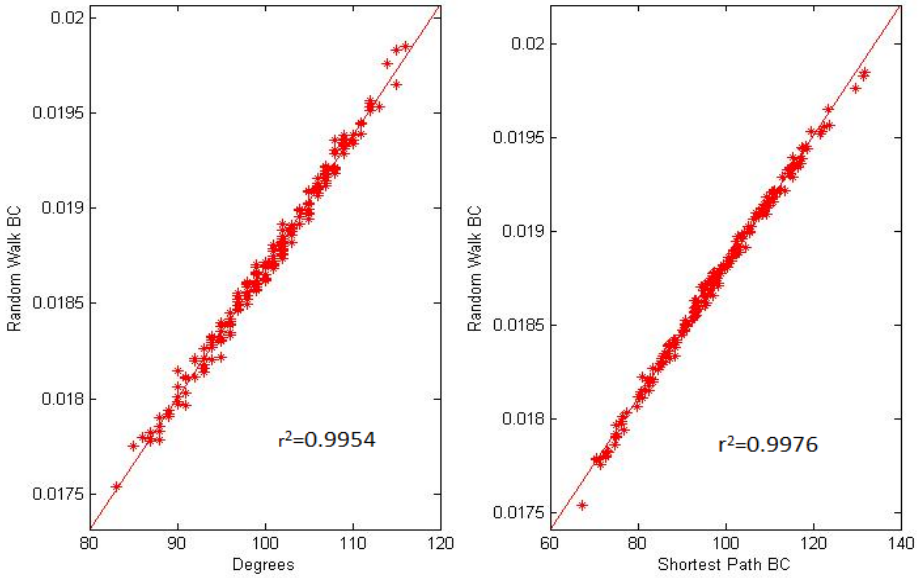


Figure 9: Plots of Correlations RWBD and RWBS for a Random Graph with p=0.5 and N=200

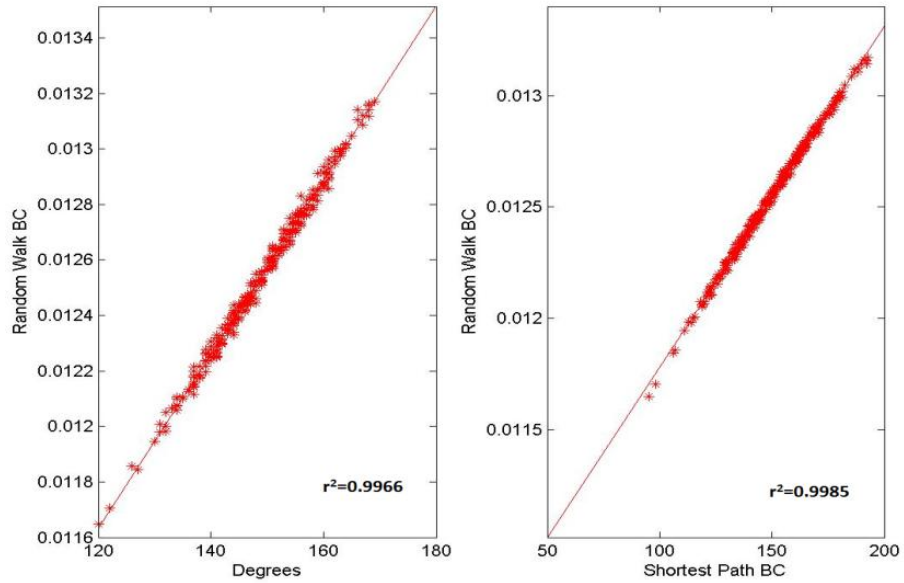


Figure 10: Plots of Correlations RWBD and RWBS for a Random Graph with p=0.5 and N=300

For each parameter set RWBD is slightly higher than RWBS. Nevertheless, we can generalize for random graphs that in random graphs random walk betweenness have very high correlation with both degree and shortest path betweenness.

In order to make sure that we reduced the effect of randomness in our measurements, we calculated correlations for average of 20 random graphs with parameters p=0.5 and N=500. Figure 11 shows the curves for these correlations. The results do not change; random walk betweenness has higher correlation with shortest path betweenness, RWBD=0.9604, whereas RWBS is very close to one, 0.9991. Furthermore, the shapes of correlation curves are linear which means that the relationships between the measures are also linear.

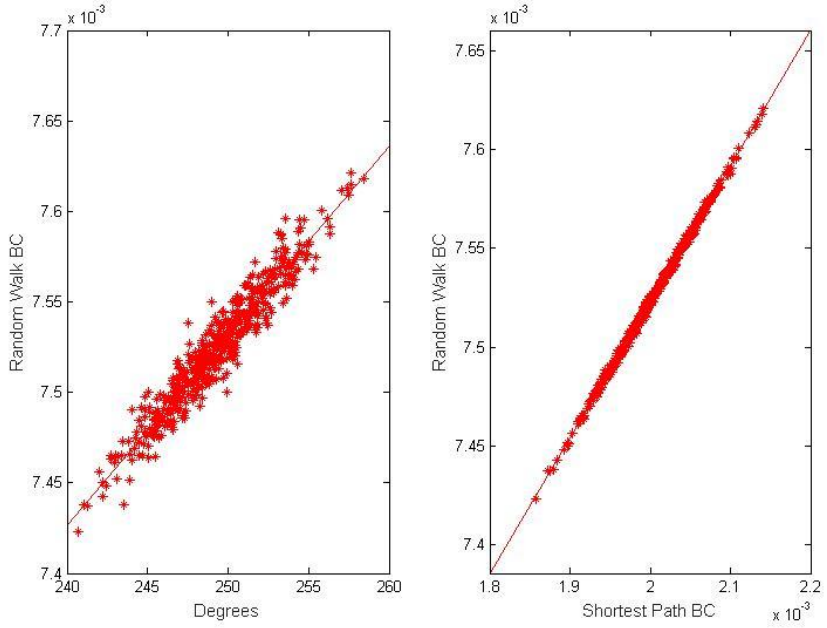


Figure 11: Plots of Correlations RWBD and RWBS for Average of 20 Random Graphs with p=0.5 and N=500

Analyses for scale-free networks resulted different than random networks as we expected. RWBD was significantly higher than RWBS for each trial this time. The results are shown in table 2.

Table 2: Correlation Coefficients of RWBD and RWBS for Scale-free Networks with Different Settings

N	*Network 1*		*Network 2*		*Network 3*		*Network 4*		*Network 5*	
	RWBD	**RWBS**	**RWBD**	**RWBS**	**RWBD**	**RWBS**	**RWBD**	**RWBS**	**RWBD**	**RWBS**
600	0.9962	0.9057	0.9972	0.9458	0.9963	0.9281	0.9960	0.9497	0.9956	0.9071
1000	0.9971	0.9076	0.9966	0.9235	0.9966	0.9229	0.9971	0.9149	0.9965	0.9259

An example of correlation plot for each parameter setting is given with figures 12 and 13. In the figures plots on the right are more chaotic than the ones on the left which visualizes the difference between correlations RWBD and RWBS.

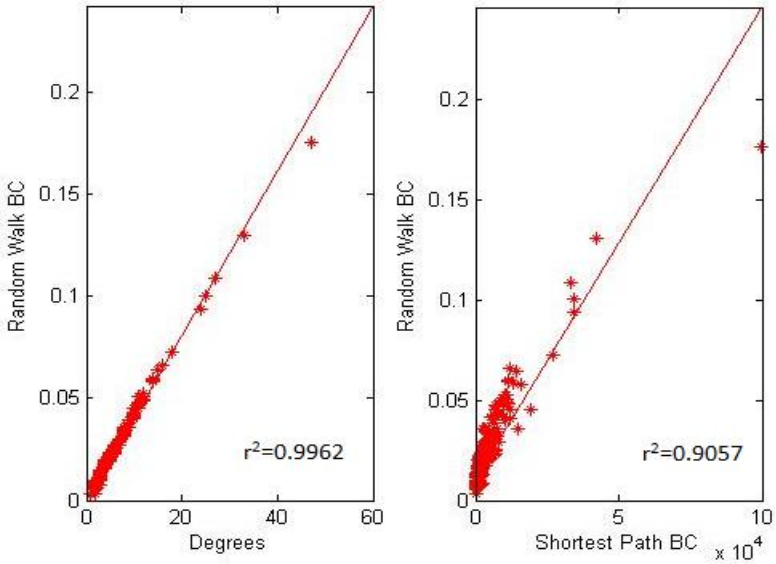


Figure 12: Plots of Correlations RWBD and RWBS for a Scale-free Network with 600 Nodes

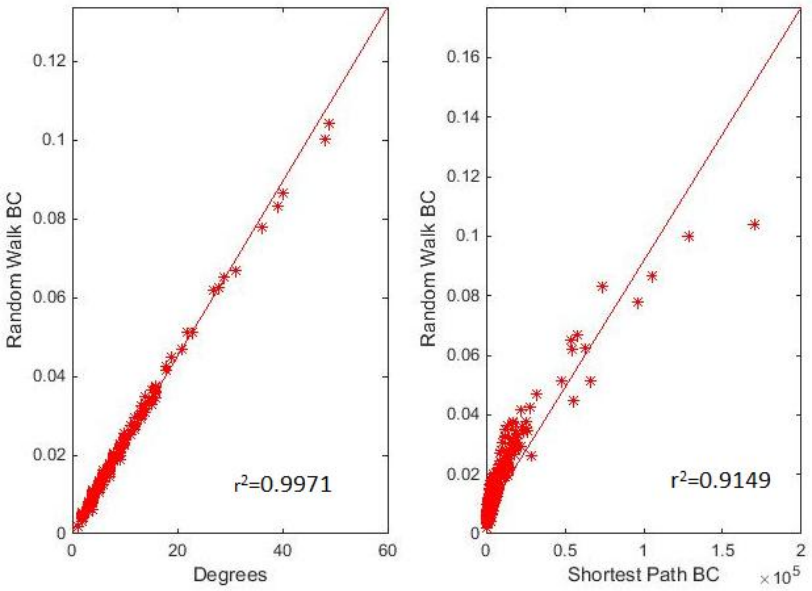


Figure 13: Plots of Correlations RWBD and RWBS for a Scale-free Network with 1000 Nodes

As we did with random graphs we also took the average correlations of 20 scale-free networks with N=600 in figure 15, again with purpose of observing them without the noise of the data. The results are not surprising once more; random walk betweenness has very high correlation with degree distribution, 0.9975, and slightly lower correlation with shortest path betweenness 0.9599.

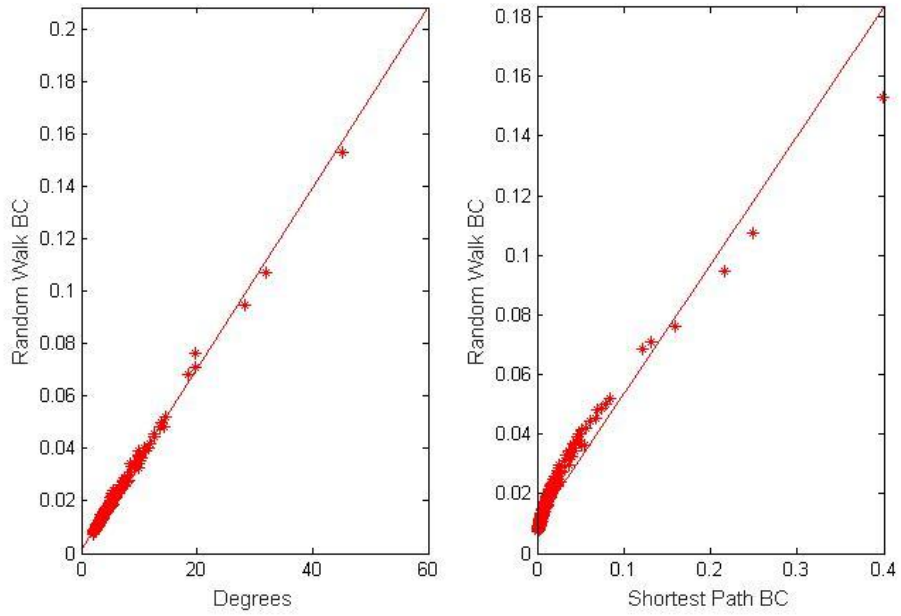


Figure 14: Plots of Correlations RWBD and RWBS for Average of 20 Scale-free Networks with 600 Nodes

Evaluation of figures 11 and 14 proves different levels of resemblances among plots in figures 1, 5 and 19 and among plots in figures 3, 4 and 20. We've already discuss these similarities for random-walk betweenness shortest-path betweenness for random-graphs and scale-free networks in section 2.1.2 where the curves are quite identical for random-graphs. Additionally, we observe that degree distribution curve of scale-free networks in figure 20 has almost the same pattern as random-walk betweenness of them in figure 4, which is less clear for random-graphs comparing figures 3 with 20.

The fact that RWBD was higher for scale-free networks is because of the hub formation in scale-free networks. Since scale free networks have hubs with very large degrees and other vertices with few degrees, this forces many random walks to pass through hubs. Thus, vertices with large degree are also those with high random walk betweenness. However, shortest paths are likely to pass through vertices with smaller degrees.

For the most of the figures in this section correlation plots show a linear correlation among the measures. This indicates that the relationships among these centrality measures are linear. Except the plot on the right hand side in figure 14, in this plot the relationship between random-walk and shortest-path betweenness for random graphs seems to have a nonlinear correlation.

2.3. Discussions

In this section we studied random walk and shortest-path betweenness for different types of networks. Especially, we compared betweenness measures of scale free networks with random graphs with different parameters. Random graphs have poisson shaped random-walk and shortest path betweenness curves. On the other hand, these curves for scale free networks are skewed, especially when we view them in logarithmic scale we notice that they have a linear curve, so they are very close to power law distribution.

Additionally, for random graphs shortest-path betweenness have greater correlation with random walk betweenness in comparison with degree centrality. On the contrary, for scale free networks correlation coefficient of RWBD is greater.

In the next chapter, we will represent our network model, and perform structural analyses on it. Conjointly, in later chapters we will benefit measures, which we reviewed here, in favor of analyzing our network model and comparing it with scale free networks.

Chapter 3

Generation of Activity Driven Networks and Understanding Their Structure

So far, we evaluated the structure of known networks using centrality measures for making sense of the network data we obtained from our measurements. Another part of this research is extending the analysis of Perra et al. (2012) [24] on network modeling. Researchers have been trying to imitate patterns of real life networks in order to understand the implications of these patterns. Modeling of static networks like scale-free networks or Erdős-Rény model has been greatly interested by network scientists for years [3] [9]. However, in real life interactions of social networks are more dynamic, since new ties are added, old ones are removed or new members are included in society expeditiously. It follows that, studying networks that evolve as the time passes, time-varying networks, meets the need of modern world significantly better [16]. World Wide Web and social networks like communication networks or contingency networks are examples of time-varying networks. In these networks time scale is separated in to small time windows and the accumulation of connections over time are captured for each discrete time step. Besides, starting with degree distribution their topological properties are analyzed on a time-integrated perspective of the system.

Perra et al. (2012) [24] offered a network model that resembles real-life social networks because of its response to centrality measures which is notably similar to centrality measures of the datasets that they examined. They studied the degree distributions of three datasets at

three time windows concerning determination of activity potential of each node. Activity potential of an agent in a given time window is given with formula below:

$$x_i = \frac{numberofinteractionsvertexi made in \Delta t}{numberoftotalinteractions made in \Delta t} \quad (2)$$

They suggested that deriving activity potentials of vertices from an activity potential distribution function F(x), which is either a known probability distribution or chosen from an empirical data. In addition, they fixed the bounds of this function to 1 and ϵ; so that, it is possible to avoid from divergences of the distribution close to 0.

Using this inside now we will illustrate generation of activity-driven networks.

3.1. Activity-Driven Network Generation

Parameters that we need to set before explaining the generation process are activity potential distribution function F(x), fixing parameter for activity potentials η, and number of links that each chosen node will form m.

Procedure:

1) For the setup procedure:
   - First, we create N disconnected nodes
   - Next, we assign activity potential x_i to each node, from the selected activity potential distribution and find activity rates $a_i = \eta x_i$

2) At each iteration or each time window Δt a vertex i becomes active with probability $a_i \Delta t$ and ties m edges to others. We perform the actions below to complete this task:
   - For each node
     - First, we generate a uniform random number
     - Next, we set the node active if its activity potential is higher than the chosen random value.

- Let each active node create m links with uniform randomly chosen others.(active or passive)

3) At the next time window network forgets all the edges created before and we go back to step 2.

In figure 15 red nodes are the active ones for that time period.Parameter m is determined as 2; thus, each active node connects to 2 other nodes.

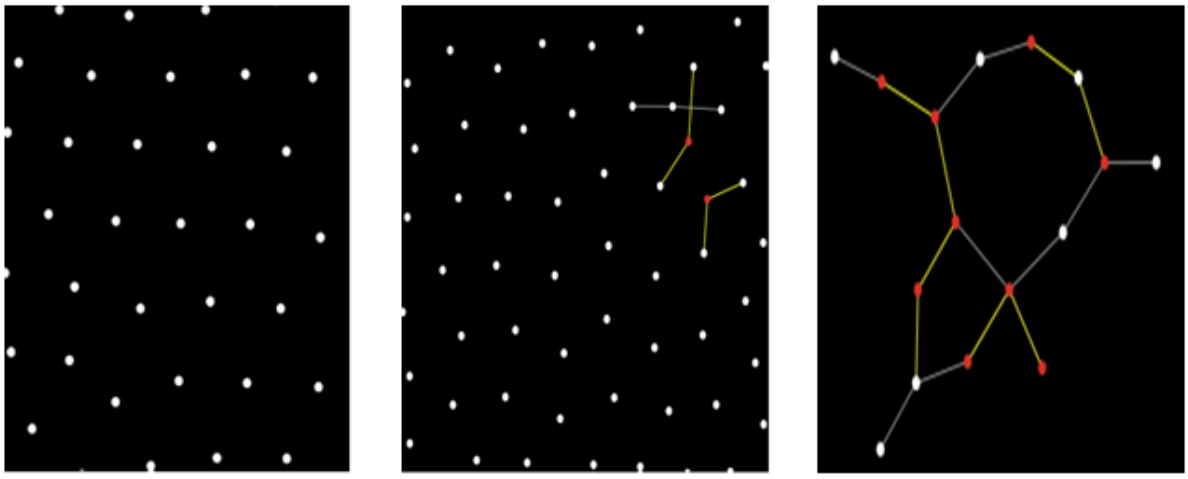


Figure 14: Simulation of Activity Driven Network Generation

We operated with NetLogo software to code this algorithm, and we performed most of the study with Power Law distribution as activity potential distribution. Nonetheless, we also generate activity-driven networks with uniform distribution as activity potential distribution and compare it with activity driven network generated from power-law distribution. Firstly, we will examine Power-Law distribution deeper to understand its affects on activity driven networks better.

3.2. Power-Law Distribution

A *power law* displays a functional relationship between two magnitudes, where one decreases as a power of another. Power laws can be shown with the general expression $f(k) = ak^{-\propto}$. If we take the logarithms of both sides of this equation [8] we reach to the equation (3):

$$logf(k) = loga - \propto logk \quad (3)$$

If we look at this equation more carefully, and plot logf(k) as a function of logk we see a diminishing straight line with slope -$\propto$ and loga as y-intercept. Hence, this method provides a quick test to check whether a dataset displays a power-law distribution or not.

Power-laws have been discovered many times in social science, physics and biology. Especially degree distributions of some real life datasets exhibit power-law distribution features. However, empirical degree distributions often do not hold power-law for the entire range. Most of the time degree distributions have a long tail as in power-law distribution; thus they follow power-law distribution for large values of k, but not for small-k regime.

Power-laws have mean only if the exponent $\propto$ is in range $2 < \propto < 3$, and almost all real life examples have an exponent in this range [19]. On the other hand, traditional variance definition does not succeed for power laws. To illustrate this one should think about a distribution that shows total wealth of people living in USA. If the richest man has 20 Billion US dollars is the variance of this dataset meaningful in comparison to average wealth?

Now, let us investigate the underlying process of power-law with the following procedure as Newman (2010) [19] states:

1. Nodes are created one by one, starting with node 1, until N^{th} node.
2. After node j is created, it ties one node among the earlier created nodes with some probability p or with probability (1-p) it will choose a page i to connect with probability proportional to its current number of neighbors.

In this algorithm nodes copy connection behavior of earlier nodes. If p is smaller than the copying behavior is more common, which causes the exponent $\propto$ to be smaller; thus, popular nodes become even more populated and it is a lot more likely to observe nodes with extremely high degrees. Consequently, it is created by rich-get-richer rule and known as *preferential attachment* phenomena.

We used a power-law distribution with $\propto$=3 and we set the lower bound as $\in = 10^{-3}$ in our analyses.

This algorithm introduces one of the several methods of generating a scale free network.

Degree distribution of scale free networks has power-law distribution.

3.3. Mathematics of Activity Driven Networks

An integrated network is defined as union of all the networks generated in the previous time windows. However, we consider only homogeneous network, links do not have strengths and we do not allow multiple edges. We should follow the steps of Perra et. al. (2012) [24] below, in order to compute the degree distribution of an activity driven network.

At a time window Δt average number of active nodes:

$$N_t = N\langle a \rangle = N\eta\langle x \rangle \tag{4}$$

Since every active node creates m links, average number of links per time window will be:

$$E_{t=} mN\eta\langle x \rangle \tag{5}$$

Hence, average degree per unit time:

$$\langle k \rangle_t = \frac{2E_t}{N} = 2m\eta\langle x \rangle \tag{6}$$

Now, we shall consider the integrated network, this will be more complex since not only each node forms m links when it's active, both also it can receive some links from other active nodes anytime during the simulation. Let us call first type of degree out-degree, and the second one in-degree [24]. If we run the model for a time interval T, then a vertex will try to send on average $\mathrm{Tma_i}$ links. However, some of these links will not contribute to out-degree since they have been connected before. We can formulate this as Polya's urns problem in which there are N different balls. We will make $\mathrm{Tma_i}$ extractions and observe the number of different balls which represents number of different links a node can form. The probability of extracting d balls is given by equation (7):

$$P(d) = \binom{N}{d} p^d (1-p)^{(N-d)} \tag{7}$$

And since each ball will be returned to the urn:

$$p = 1 - \left(1 - \frac{1}{N}\right)^{Tma_i} \tag{8}$$

for large N $\left(1-\frac{1}{N}\right)^{Tma_i}$ approaches to exp $(-mTa_i/N)$.Proof of this can be done by taking natural logarithm(In) of both side and then the limitation for N goes to infinity as shown with equations (9) and (10). Limitation is performed by using L'hospital rule.

$$mTa_i \ln\left(1-\frac{1}{N}\right) = \frac{-mTa_i}{N} \tag{9}$$

$$\lim_{N\to\infty}\left(\frac{\ln\left(1-\frac{1}{N}\right)}{-\frac{1}{N}}\right) = 1 \tag{10}$$

Mean of binomial distribution is np hence average out-degree of an integrated network is:

$$k_T^{out}(i) = N(1-e^{-mTa_i/N}) \tag{11}$$

Next step is finding in degree, average degree that a node received from other active nodes. We need to know number of nodes that have not received any link from vertex i. The probability that a node has not picked up any links from I until time T is given with following equation :

$$\left(1-\frac{1}{N}\right)^{mTa_i} = \exp\left(-mTa_i/N\right) \tag{12}$$

Hence the average number of nodes that has became active at least once, but that has not received any connection from vertex i is TN⟨a⟩exp $(-mTa_i/N)$. Each of these nodes will hit vertex I with a probability of $\frac{\mathrm{m}}{\mathrm{N}}$. Hence, the number of links that vertex I will receive is$\mathrm{mT\eta\langle x\rangle exp\,(-mTa_i/N)}$. The degree of vertex i is sum of its in and out degrees:

$$k_T(i) = k_T^{in}(i) + k_T^{out}(i) = N\left(1-e^{-\frac{mTa_i}{N}}\right) + \mathrm{mT\eta\langle x\rangle}\exp\left(-\frac{\mathrm{mTa_i}}{\mathrm{N}}\right) \tag{13}$$

$$\sim \mathrm{N}\left(1-e^{-\frac{mTa_i}{N}}\right) = \mathrm{N}\left(1-e^{-\frac{mT\eta x_i}{N}}\right) \tag{14}$$

If we extract x from this equation (14) and write it as a function of degree:

$$x(k) = \frac{N}{\eta m T} \ln\left(1 - \frac{k}{N}\right) \tag{15}$$

At this point $P_T(k)dk \sim F(x)dx$ or more precisely as in equation (16):

$$P_T(k)dk \sim F\big(x(k)\big)\frac{dx(k)}{dk} = \frac{1}{\eta m T}\frac{1}{1-\frac{k}{N}} \mathrm{F}\left[-\frac{N}{\eta m T}\ln\left(1-\frac{k}{N}\right)\right] \tag{16}$$

Consequently for the limit of small k/N we can reach the approximation:

$$P_T(k) \sim \frac{1}{\eta m T} \mathrm{F}\left[\frac{N}{\eta m T}\right] \tag{17}$$

Equation (17) shows the relationship of activity potential function with degree distribution of the network.

3.4. Comparison with the Original Work

Our purpose was to reproduce the work of Perra et. Al.(2012) [24], and in extend to that analyze this model with various calibrations. Therefore, we would like to attain their effort correctly. For this goal, we compare their resulting degree distributions with ours in figures 16, 17 and 18. All parameters except the exponent of power-law distribution is the same in both cases N=5000, m=2, η=10 and $\epsilon \leq x \leq 1$ with $\epsilon = 10^{-3}$. Yet, we were limited to perform our analysis a power-law distribution with $\propto = 3$, although they implemented $\propto = 2.8$ as activity potential distribution. The reason that we chose a different exponent is related with the obstructions of NetLogo software. Mathematical computations such as $k^{-\propto}$ are not possible to pursuit with NetLogo; thus, we solved this problem by taking advantage of its preferential attachment model. This model builds scale-free networks using Barabasi-Albert algorithm [4] which works with preferential attachment model as we explained above in section 3.2. We obtain the degree distributions by averaging over 20 simulations. Figure 16 shows the degree distribution of simulations after one time step, at T=1. Figure 17 is degree distribution of the snapshot of the integrated network at T=10, and figure 18 is degree distribution at T=20. In each figure the series with black line and red-circle marker represent the results of Perra et. Al. (2012) [24], and other series display our findings.

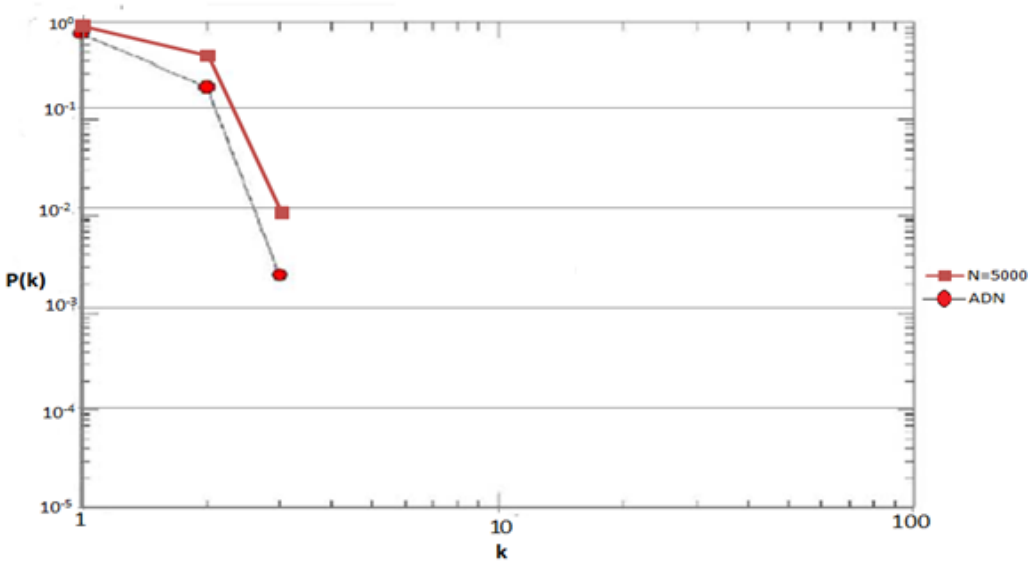


Figure 15: Degree Distributions Comparison of ADN of Perra et.al. (2012) [24] and Our Model at T=1

Although for all plots axes are logarithmic, we do not detect a straight line log-log degree distribution. Instead, we notice a skewed plot for degree distribution of activity driven network. This resolves that, activity driven network behaves different than scale-free networks even though it has a power-law activity potential distribution. In scale free networks hubs are constructed thanks to their positional advantage, and they are connected by more and more nodes passively. On the other hand, for activity-driven network hubs are created because they have greater activity rates than other nodes, hence they are enthusiastic about forming new ties.

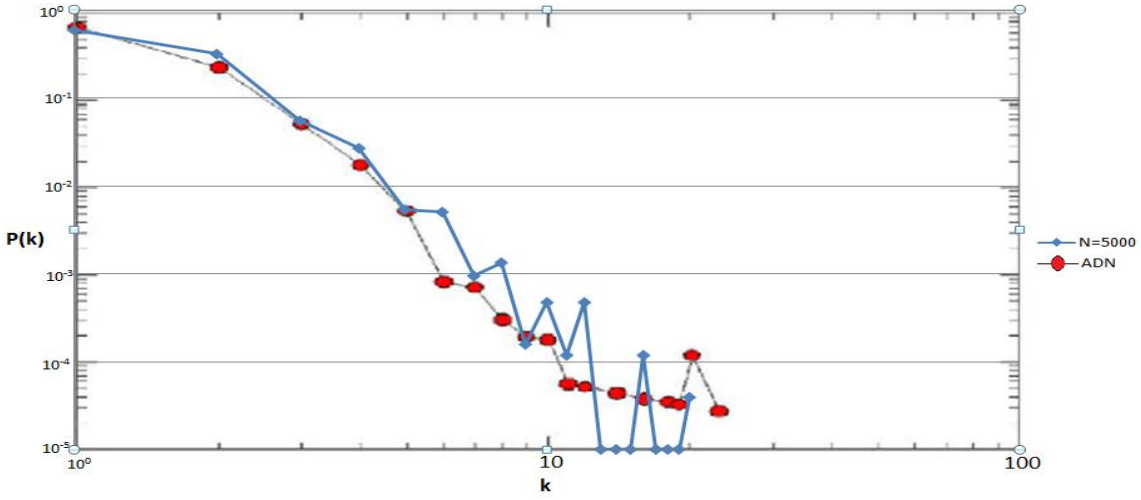


Figure 16: Degree Distributions Comparison of ADN of Perra et.al.(2012) [24] and Our Model at T=10

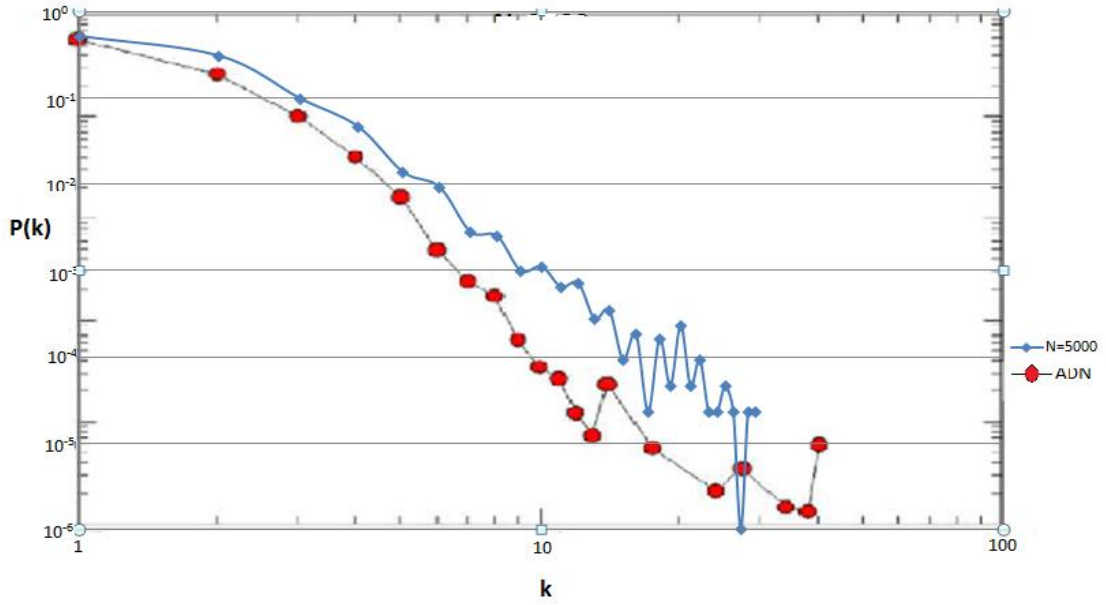


Figure 17: Degree Distributions Comparison of ADN of Perra et.al. (2012) [24] and Our Model at T=20

Increasing the exponent ∝ from 2.8 to 3 causes a broader degree distribution for the network. In consequence, in all three figures plots of Perra et. al. (2012) [24] are below ours.

3.5. Centrality Measures for Activity Driven Networks

In this section we will probe activity driven networks (we shall refer this as ADN) by adopting centrality measures and clustering coefficient on it.

3.5.1. Degree Distribution

Degree distribution is the most fundamental and straightforward centrality measure discovered. Therefore, initially we will study degree centrality of nodes in ADN. The degree of a vertex is known as the number of links it has. Additionally, for social networks the degree of a person can be defined as the number of friends he has [1] [6].

We present degree distributions of random graphs and scale-free networks consecutively in figures 19 and 20.

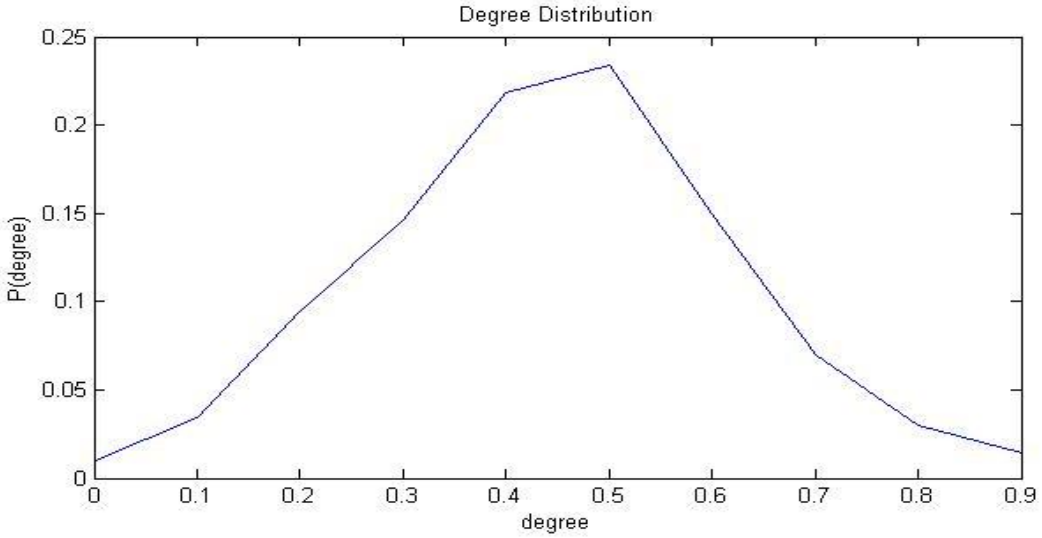


Figure 19: Average Degree Distribution of 20 Random Graphs

Figure 19 displays average degree distribution of 20 random graphs with N=500 and p=0.4, and in figure 20 we observe the average degree centrality of 20 scale-free networks with 600 nodes. Plus, in figure 20 both axes are logarithmic scaled.

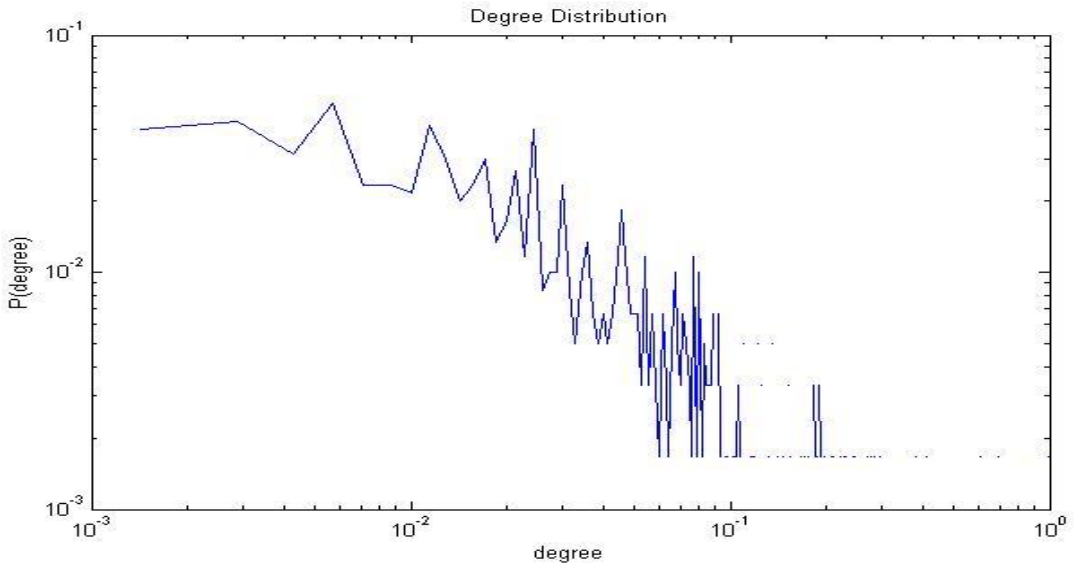


Figure 18: Average Degree Distribution of 20 Scale Free Networks

Additionally, degree distribution of ADN for different time windows are given in figures 17 to 18. Although we use power-law distribution as activity potential distribution for our analyses, it is clear also from figures 16,17 and 18 that the degree distribution of ADN in logarithmic scale does not possess a linear curve; thus, it does not behave exactly like scale free networks.

We simulated the model for 20 times and obtained degree distribution in logarithmic scale for N=300, and m=2, η=10 and $\epsilon \leq x \leq 1$ with $\epsilon = 10^{-3}$ at time windows T=25, T=50, T=75 and T=100. Our choice of number of nodes represents the population of the community; hence, the bigger N gets the closer our simulation is to real life dataset. However, if we choose to study with a large network, it is extremely time consuming to perform all the analysis on the network and unnecessary for scope of this project. Other parameters demonstrate how fast the network grows, and they also determine average number of active nodes and number of links connected per time window. For example, in our settings we used the settings given below with equations (18), (19) and (20).

- Average number of active nodes per time window:

$$N\langle a\rangle = 300 \times 0.00999 \cong 3 \tag{18}$$

- Average number of links per time window:

$$mN\langle a\rangle = 6 \tag{19}$$

Hence,

- Average degree:

$$2m\langle a\rangle \cong 0.04 \tag{20}$$

Using the same parameters we will also examine clustering coefficient and average path length distributions of ADN.

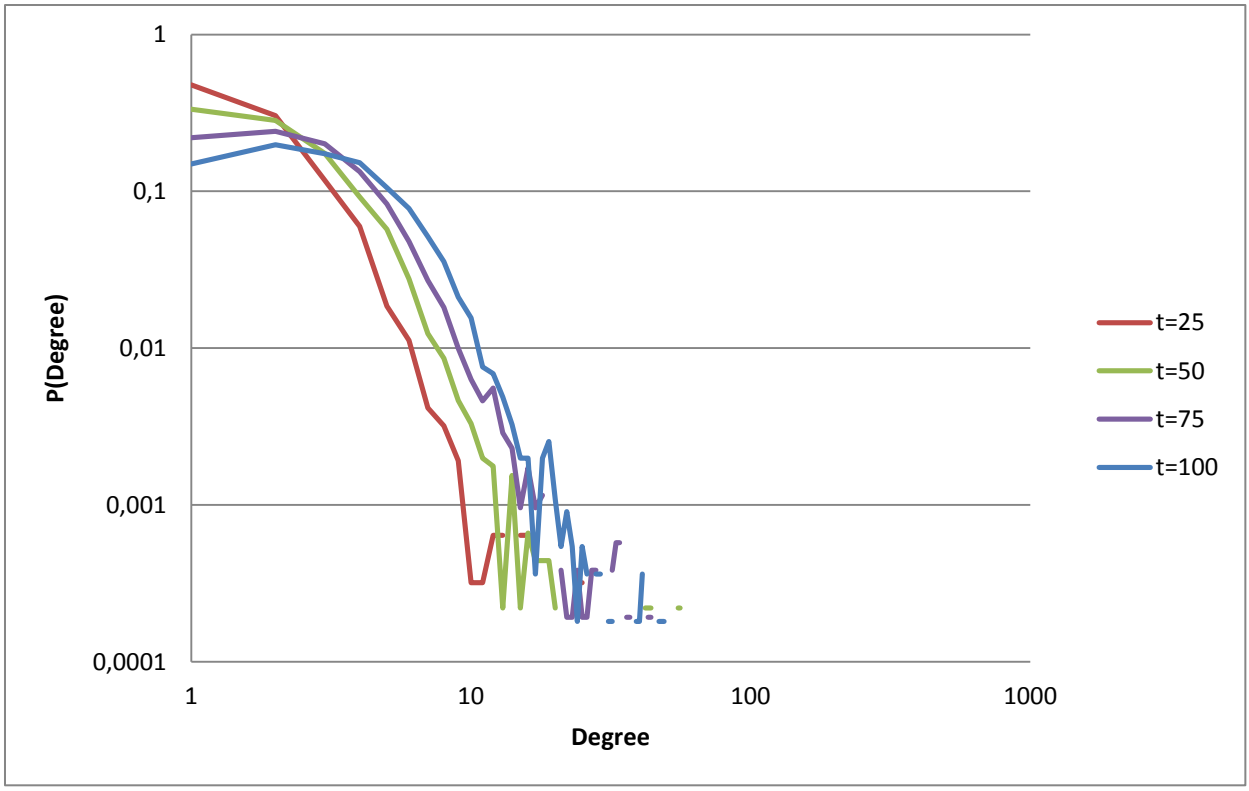


Figure 21: Degree Distribution of ADN at Different Time Windows and in Logarithmic Scale

In figure 21 the effect of time window on degree distribution of the integrated network can be identified distinctly. The plots are highly left skewed, so the network contains extreme degree values. As the time passes and number of connections boost in the integrated network, the plots become broader and they approach to Gaussian form.

We practiced goodness of fit test on our data concerning whether degree distribution of ADN fits to any random distribution with Arena Input analyzer. Input analyzer did not obtain any good fit; for all time windows p value was lower than 0.005. The best fits that the program offered was for ADN at t=100 beta distribution with square error 0.0420, for ADN at T=75 to Erlang distribution with SE=0.001292. Moreover ADN at T=50 was fitted to gamma distribution best and ADN at T=25 with square error 0.000888 was closest to Exponential distribution with SE=0.000945. These tests certify that degree distributions of AD at given time windows do not fit to any known random distribution. Since log-log degree distributions do not have a straight shrinking curve they are not power-law either.

3.5.2. Clustering coefficient

Clustering coefficient of a node A is the fraction of pairs of A's neighboring nodes that are linked to each other [8] [13] [27]. More specifically:

$$\frac{number of pairs of A^{'}s neighbors that are connected}{Binary combination of number of A^{'}s neighbors} \tag{21}$$

Clustering coefficient of a network is the average clustering coefficient of its nodes. It reveals the connectedness of a network. Thus, clustering coefficient of a fully connected network is 1. Moreover, clustering coefficient of a scale free network is acutely low compared to clustering coefficient of random graphs. Clustering coefficient (we will refer it as CC) of random graphs is fraction of average degree over number of nodes $\frac{\langle \boldsymbol{k} \rangle}{\boldsymbol{N}}$ [27], which is around 0.5, whereas this number for scale free networks is much smaller, and it is around 0.01-0.02. Nevertheless, clustering coefficient of preferential attachment model is zero, since it never allows triadic closures.

In figure 22 we remark clustering coefficient distribution of random graph with p=0.5 and N=500.We took the average of 20 graphs for the analysis. The curve has a poisson shape just like its degree distribution. For the averaged random graph average degree $\langle \boldsymbol{k} \rangle$ is 249.3048 and there are 500 nodes; thus we expect clustering coefficient of the network to be around 0.4986, and the calculated clustering coefficient is 0.4996. Hence, estimation of Watts et. al. (1998) [27] is very close to real CC with the error 0.1 percent.

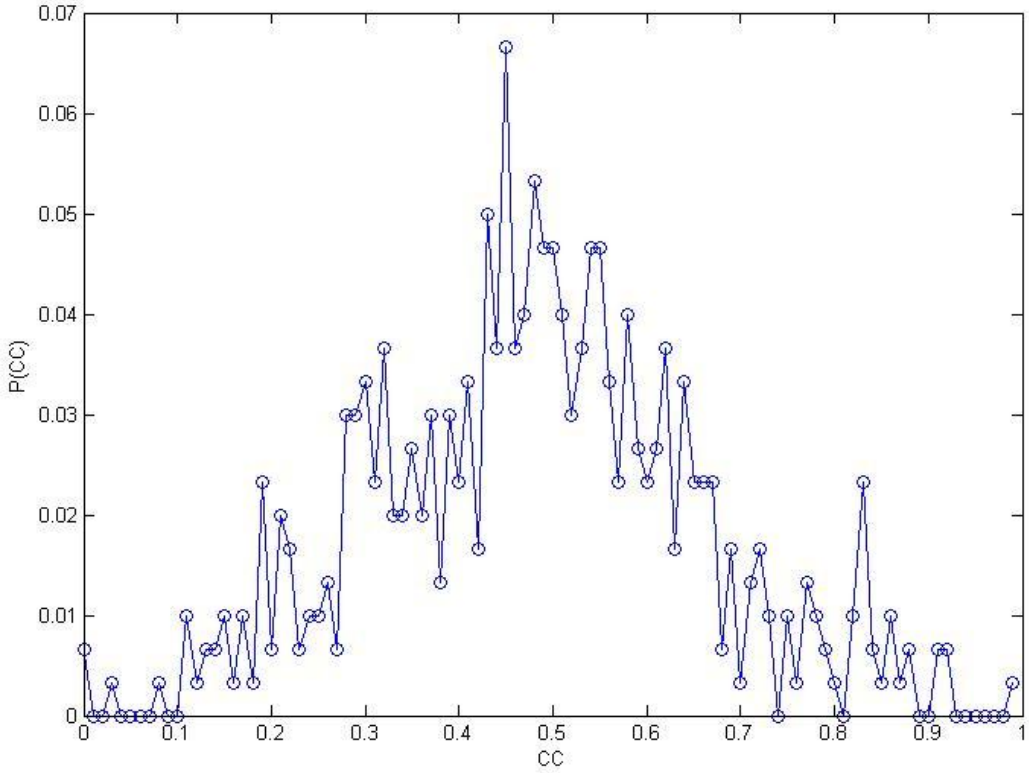


Figure 22: Clustering Coefficient Distribution of Random Graph

On the other hand, clustering coefficient distribution of a scale-free network is hardly visible because of overly low values. Clustering coefficient of averaged scale free network is 0.0282, which empirically reveals that clustering coefficient of scale free networks is significantly lower than clustering coefficient of random graphs.

Clustering coefficient distribution of activity driven network functions like that of scale free network. It is also quite low for each time window we tested.

Average clustering coefficient of the whole network in different time windows can be inspected in figure 23. Especially for small T values, the clustering coefficient is immensely small as for scale-free networks, and it increases linearly as the time passes.

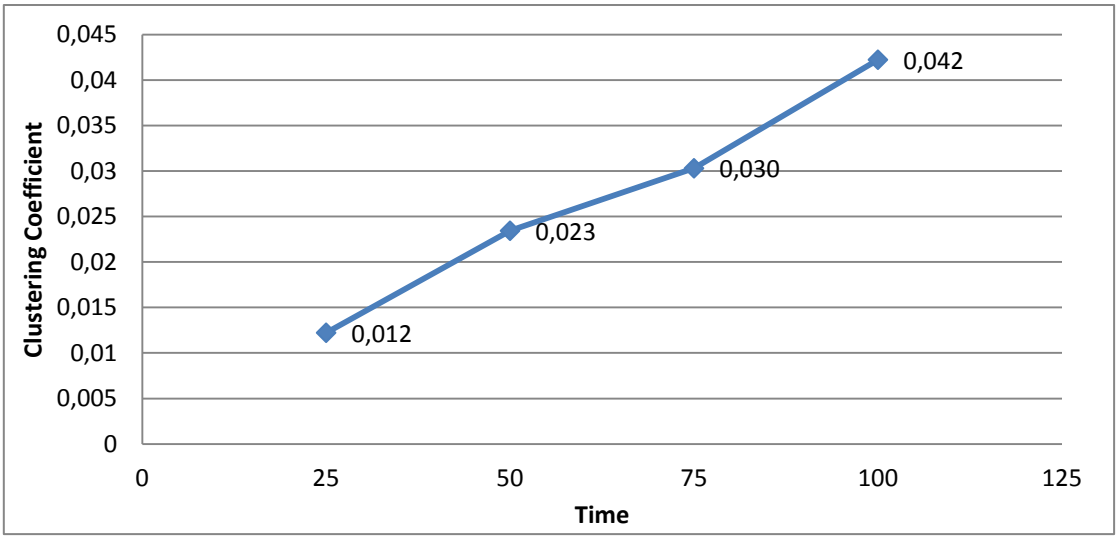


Figure 23: Average Clustering Coefficient of ADN for Different Time Windows

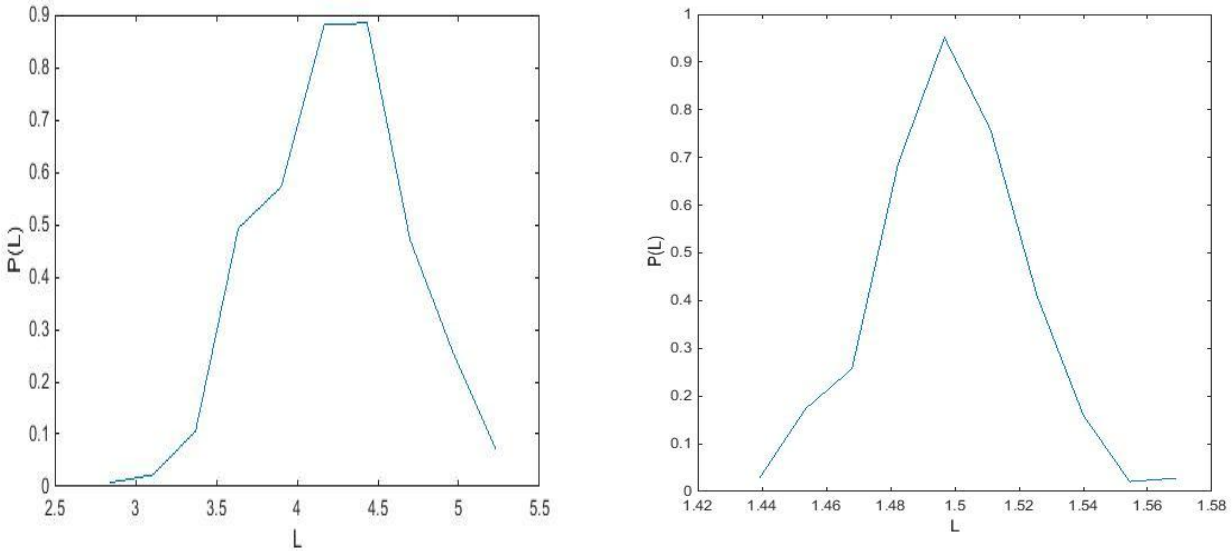
3.5.3. Average Path Length

Consider a vertex A, let us calculate the average shortest path distances from A to all other vertices in the network. This measure is called path length L. If we perform the same analyze for each vertex in a network, and obtain the average we attain average path length of the network. Average path length quantifies the efficiency of information or mass flow on a network.

However, it is altered from shortest path betweenness inasmuch as shortest path betweenness of a node computes the fraction of times that it is located in the shortest path from a source node s to a target node t, averaged over all pairs of s and t.

Average path length for random graphs is around $\frac{\ln N}{\ln\langle k\rangle}$, and for scale-free networks it is numerous times higher. This means that in random graphs each node is easier to reach, than the nodes in scale-free networks.

Average path length distributions of a random graph and a scale-free network are given in figure 24 (a) and (b). L for figure 24-A is 4.00118, and for B is. 1.4974. For scale free network average closeness centrality is almost 3 times greater than random graphs.



(a) **(b)**

Figure 19 : Average Path Length Distribution for Different Types of Networks. On the left L for a Scale-free Network, and on the right for a Random Graph

Sometimes a network is not connected, especially for evolving graphs this is very much likely for the earlier ages of a network. In this case path lengths for some pairs of nodes will be infinity, so the average path length of the network. To solve this issue we will use 1/L. Thus, if a node is not connected to anywhere its 1/L value will approach to zero and if as its reachability is high from other nodes along the shortest paths; then, 1/L value will be close to 1. When we look at the graphs of l and 1/L we recognize that they are symmetrical as in figure 25. On the right side we detect average path length distribution for a random graph with red line, and the plot with blue color on the left side shows 1/L distribution of the network.

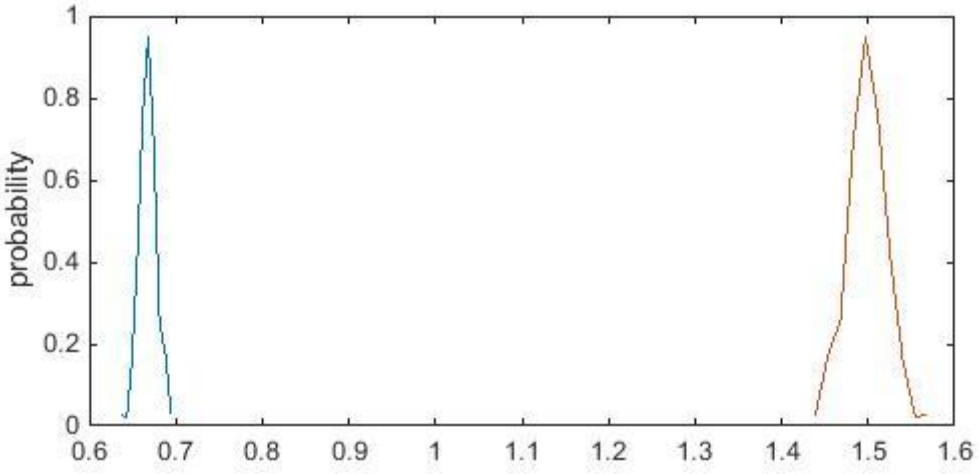


Figure 25: 1/L Distribution of a Random Graph is Blue Curve and Average Path Length Distribution is Red Curve

Behavior of activity driven networks for average path length is more chaotic than both random and scale-free networks. Figures below represents 1/L data after averaging over 20 simulations.

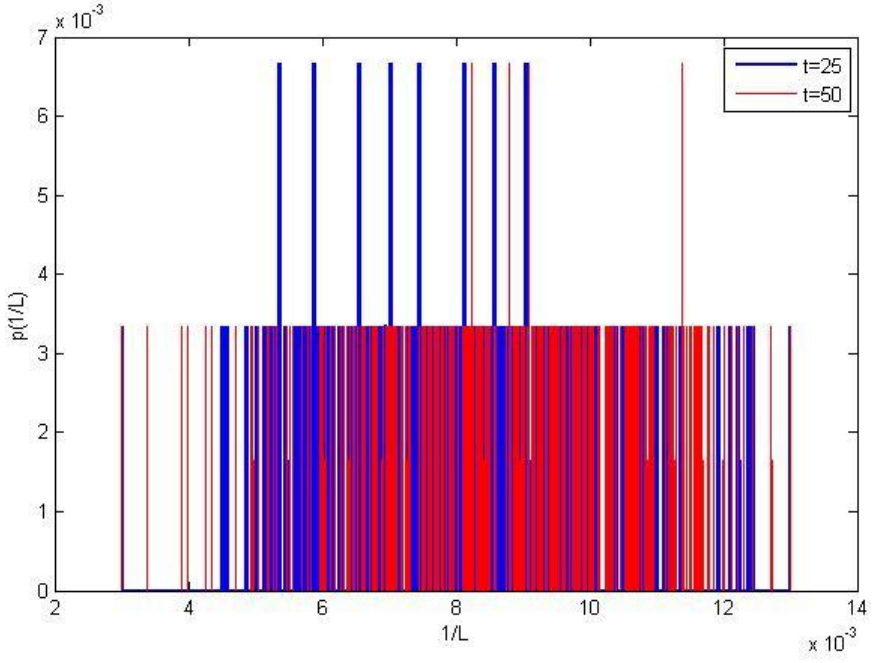


Figure 26: 1/L Distribution of ADN at Time Windows T=25 and T=50

As we observe in figure 26, 1/L distribution of ADN appears to not to acquire a coherent pattern because of the gaps in the curves. The gaps occur because the network is not fully

connected at these time windows, so L is infinity for many nodes and 1/L is very close to zero. A power-law activity driven network with 300 nodes becomes complete at around 7400 iterations with our settings. We can find approximate minimum time length to gain a fully connected network as follows; if each node is connected to one central vertex there will be 299 links in the network. Since the model forms 6 links per time window on average, lower bound is 50 iterations to have a fully connected network. Considering extremeness of our assumptions, it is quite fair that we do not have a fully connected network even at T=100. If there is one disconnected node then the path length of all the other nodes will be infinity, which is why our analysis on average path length for ADN including all nodes not a healthy analysis.

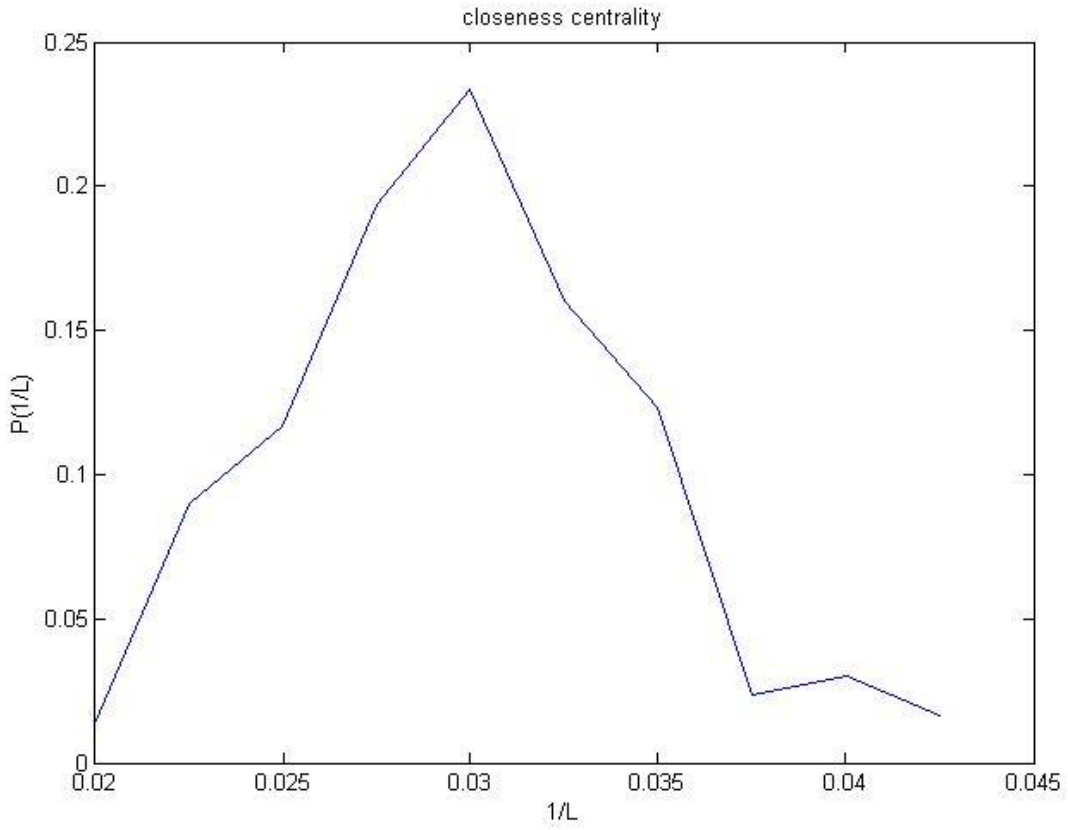


Figure 27: 1/L Distribution for ADN at Time Window T=300

To solve this problem and study closeness centrality of ADN we calculated average path length after ADN becomes fully connected. In figure 27 given 1/L distribution of ADN at T=300. This time we observe a slightly left skewed, connected curve. Average path length for ADN at T=300 is 2.450; thus, the average distance between two nodes is 2-3 links.

3.6. Activity Driven Network with Uniform Distributed Activity Potential Function

We generated activity driven networks also with uniform distribution as activity potential distribution (in order to avoid confusions we will refer to this network uniform ADN, and the earlier one as power-law ADN or only ADN). According to this, each node has a float activity potential rate $x_i$between 0 and 1, and the scaling factor η is chosen as 0.1 to decrease number of edges per iteration.

Further in this section we will investigate centrality measures for uniform ADN. Measures are calculated only for time window T=50, since it will be adequate for our purpose which is to compare centrality curves of uniform ADN with power-law ADN, and their mechanism of triadic closures.

3.6.1. Degree Distribution of Uniform ADN

The degree distribution of uniform ADN as averaged over for 20 networks with parameters N=300 and m=2 at time window T=50 is given in figure 28. Average number of active nodes in a time window for these settings is 30, so average number of links is 60 and average degree of a node is 0.4.

Average activity potential function of 20 uniform ADN is normal distribution since the sum of uniform random numbers forms normal distribution according to central limit theorem. Moreover, shape of activity potential function affects the shape of the degree distribution for ADN. Although the curve in figure 28 is slightly skewed to left, this is the reason for its poisson shape.

Additionally, random graphs also have passion shaped degree distributions. Therefore, with this setting ADN resembles random graphs more.

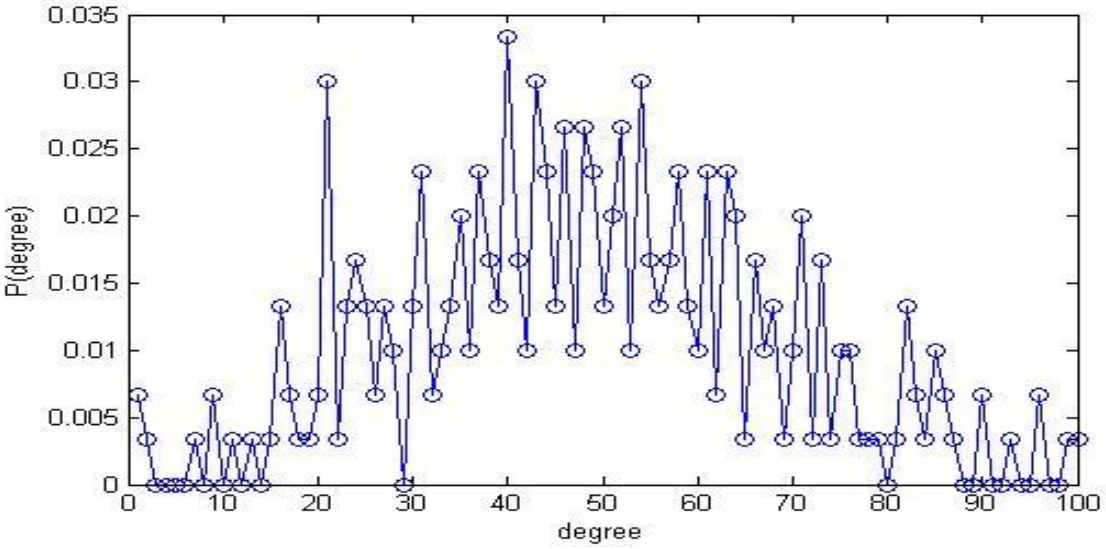


Figure 28: Degree Distribution of Uniform ADN at T=50

3.6.2. Clustering Coefficient of Uniform ADN

In figure 29 normalized clustering coefficient distribution of uniform ADN is given. The curve is slightly skewed to left, and average clustering coefficient of uniform ADN is 0.0425.

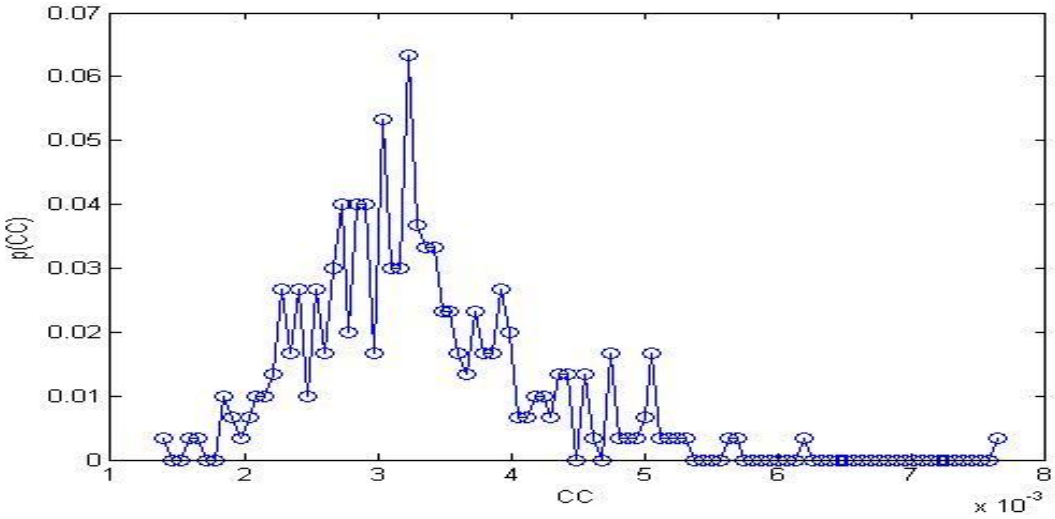


Figure 29: Clustering Coefficient Distribution of Uniform ADN at T=50

3.6.3. Average Path Length of Uniform ADN

Lastly, in figure 30 we observed average path length distribution of uniform ADN at T=50. Here we notice a quite symmetrical poisson shape curve for 1/L, and the noise in the data even though we took the average of 20 networks for our analysis.

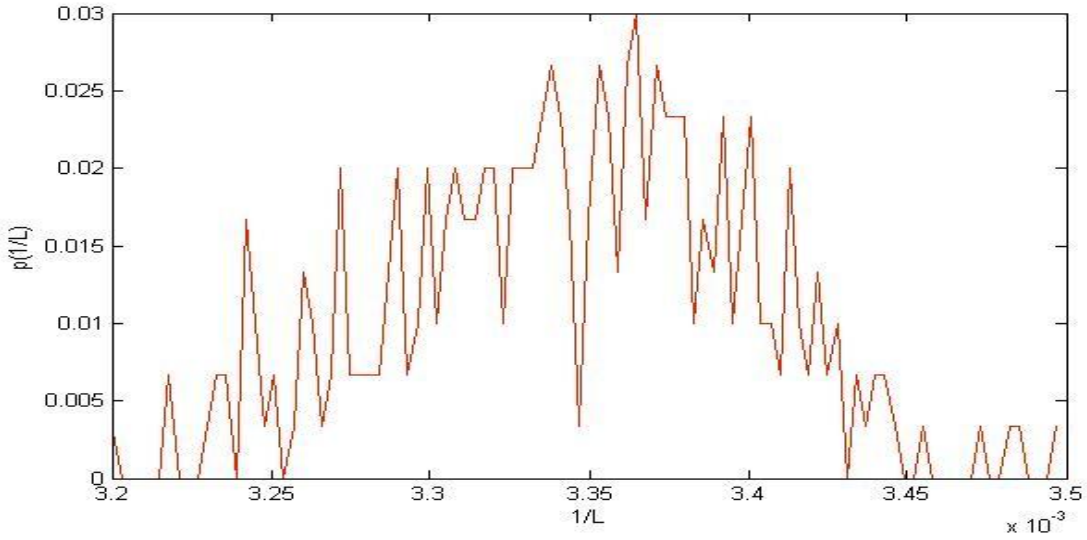


Figure 30: 1/L Distribution of Uniform ADN at T=50

When we change the activity potential structure of ADN, network's all properties are adjusting accordingly, if we compare the plots for uniform ADN with plots of power-law ADN.

3.7. Discussions

Most of the network generation algorithms focus on static networks. However, in many networks that the flow is information, links are changing rapidly some of them are vanishing and many new are created. Hear we generated time-varying networks whose all connectivity patterns depend on a single function, the activity potential function. Perra et. all.(2012) [24] empirically measured activity potential of nodes in three different real life longitudinal datasets, PRL dataset, Twitter dataset and IMDb dataset, and proved connectivity pattern similarity of ADN with these networks. The model is simple and the algorithm runs in

seconds with NetLogo. It can be used to solve analytically some network problems, like contagion processes, cascading behavior or time-scale separation issues.

This model is better than preferential attachment model in representing real life social networks. Although preferential attachment model is simple to generate, its degree distribution is power law. On the other hand, Perra et. all.(2012) [24] demonstrated that the datasets that they handled do not possess power law degree distributions; but, their agents have power law activity potentials. Moreover, preferential attachment model does not contain any triadic closures, and this situation is also unlikely for real life datasets.

The fact that ADN is markovian, is unrealistic for some real life social networks. For instance, in friendship networks we always identify triadic closures as Easley and Kleinberg (2010) [8] suggests.

Chapter 4

Betweenness Centrality for Activity Driven Networks

We argued the importance of centrality measures for understanding and analyzing a network better. Likewise, we inspected the degree distribution of ADN for different time windows in chapter 3. We also observed clustering coefficient and average path length distributions for ADN. Degree distribution and clustering coefficient distribution of ADN are both skewed and have long tales on the right side of the plot; hence they follow the pattern of activity potential distribution, power law distribution. In this chapter we will focus on shortest path betweenness and random walk betweenness of ADN.

4.1. Shortest Path Betweenness

We analyze the shortest path betweenness distribution of activity driven networks at different time windows as we did with degree distribution. We exploit Matlab's efficiency in mathematical calculations for this analysis after we generated 20 network datasets at time windows t=25, t=50, t= 75 and t=100. The plots in figures 30 to 37 are obtained by taking average of these 20 model generations, with parameters N=300, m=2, η=10, $F(x) \sim k^{-\gamma}$ with γ=3 and $\epsilon \leq x \leq 1$ with $\epsilon=10^{-3}$.

In figure 32 snapshots of shortest path betweenness distribution of ADN at different time windows and in logarithmic scaled is provided. The series in this plot are parallel to the degree distribution of ADN, in a way that they are broadly skewed and gets noisier in the tale. Conceptually, this means that there are some nodes whose betweenness value are extremely high, and many others who have low betweenness. Moreover, we notice that as the time passes the series positioned higher in the figure. For example plot of t=25 is at the most bottom of all, and slightly higher we see plot of t=50 so on. This demonstrates that the gap between probability of having low betweenness values and high ones increases as the time passes.

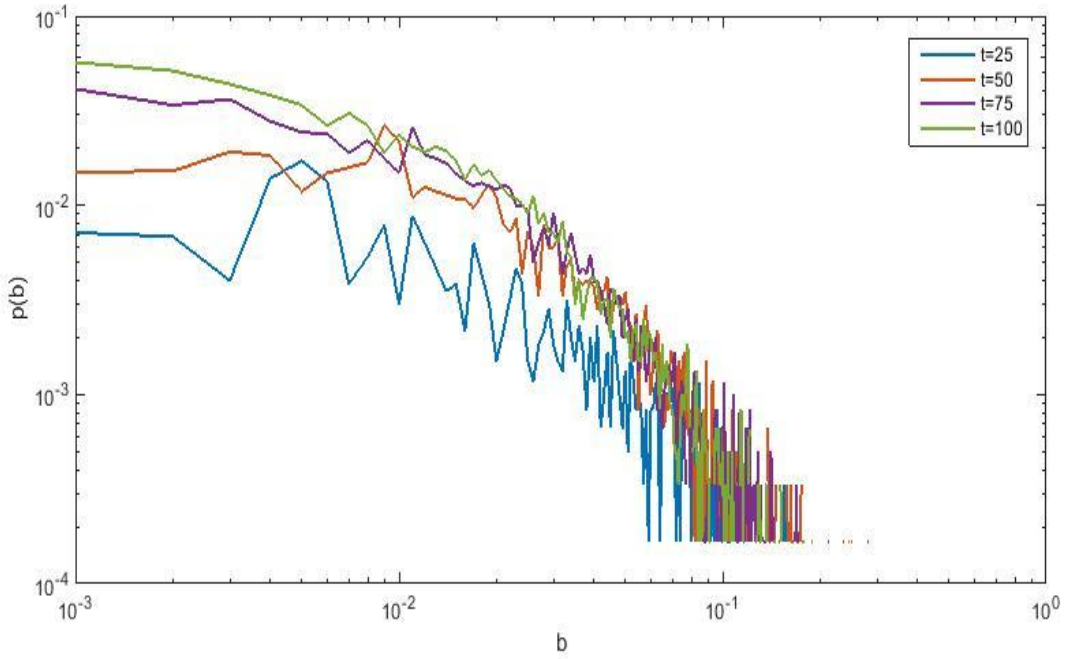


Figure 32: Shortest Path Betweenness Distribution of ADN at Different Time Windows

4.2. Random Walk Betweenness

Next, we observed random walk betweenness for activity driven networks with the same parameters as above.

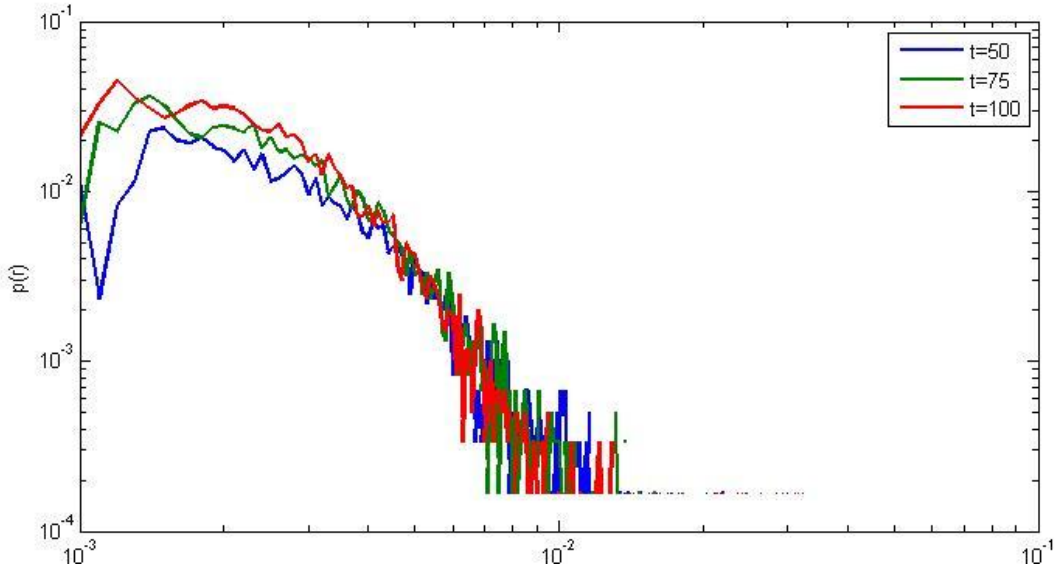


Figure 33: Random Walk Betweenness Distribution of ADN at different time windows

In figure 33 we see the effect of power-law distributed activity potential distribution more precisely for small time windows, and as the time raises activity random-walk distribution of activity driven networks resemble slightly passion distribution.

Plot of T=100 seems to be slightly above all others. However, the divergence is not as clear as in shortest path betweenness.

In the interest of understanding random walk betweenness of ADN more let us look closer to random walk betweenness distribution plot at T=100 in figure 34, which we assume is long enough for activity driven networks to show its characteristics.

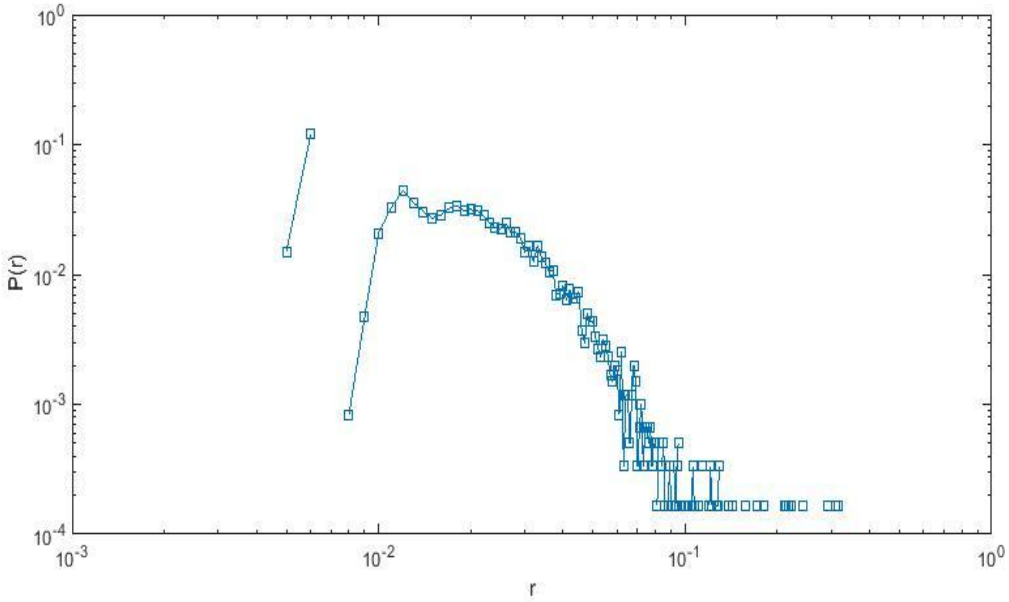


Figure 34: Random Walk Betweenness Distribution of ADN at T=100

This plot is in logarithmic scale, and it approximates neither passion nor pareto. Nevertheless, if we remove a few data points from the beginning, figure 34, again it has a largely skewed shape with a long tale as in degree distribution of ADN.

Hence, as we mentioned in power-laws section (see page 21), power-laws do not exist in real life. Only some parts of resemble power-laws, and deviate from it partially. In the case of figure 34, the data not only diverges from power-law at the tale of the curve but also at the beginning. When we remove very small values of random walk betweenness from figure 34 we obtain figure 35.

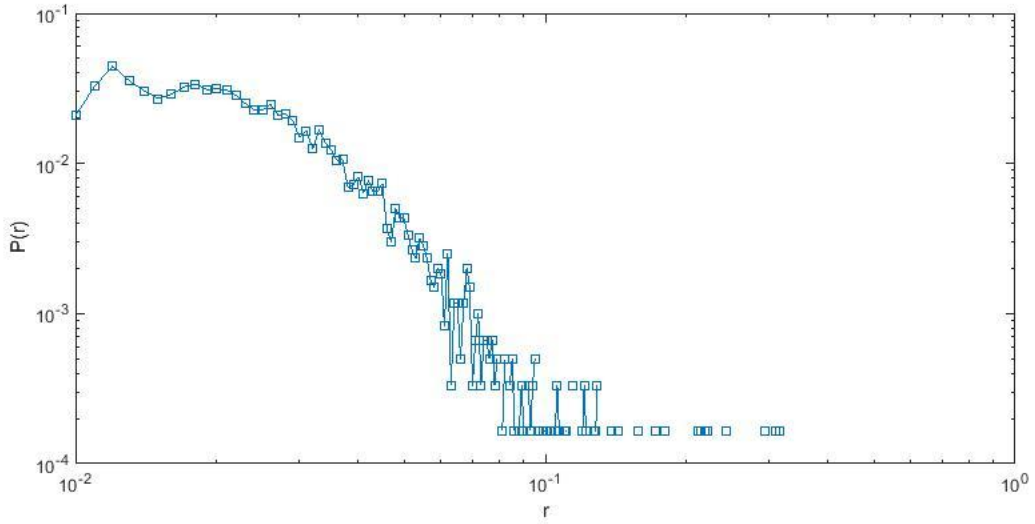


Figure 35: Random Walk Betweenness Distribution of ADN at t=100

The curve in figure 33 appears like the degree distribution or the shortest path betweenness distribution of ADN.

In conclusion, betweenness measures of activity driven networks are influenced from its activity potential distribution.

4.3. Correlations

In the earlier chapter we looked into correlations among centrality measures for several known datasets as well as scale-free networks and random graphs. Moreover, we learned that if RWBS correlation of a network is high, it is highly clustered, as in the case of random graphs. Thus, RWBS gives information about how clustered a network model is, while RWBD gives information about hub formations. This information will be helpful for as in the next chapter, while working on triadic closures in ADN.

In chapter 2 we showed that in scale-free networks the nodes which are regularly visited during a random walk also the nodes with high degrees, in other words they are hubs. However, it is less frequent that we observe nodes with high random walk betweenness are also those who stand in many shortest paths between all node-pairs. Hence, scale-free

networks have low clustering-coefficient and triadic closures do not happen frequently for them.

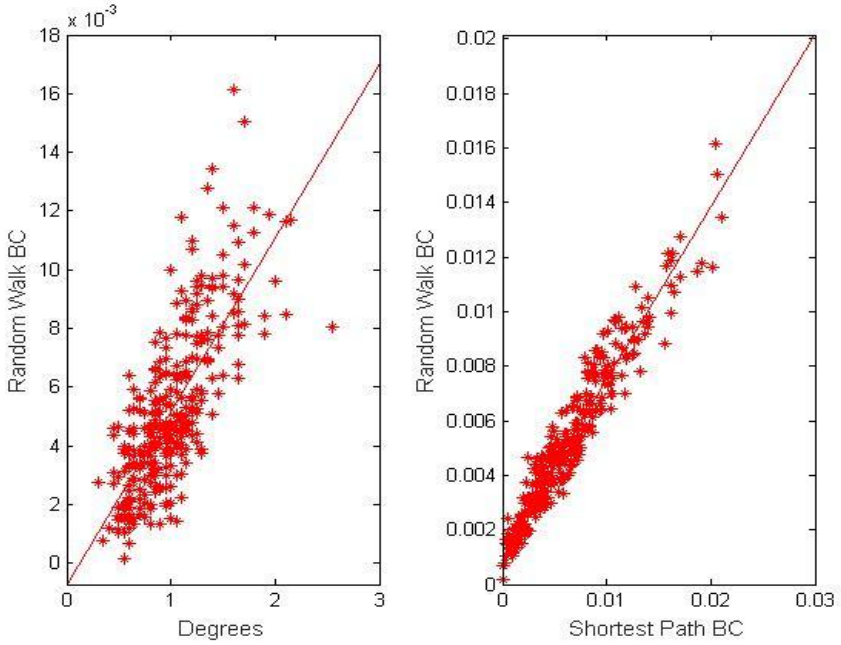


Figure 36: Plots of ADN for Correlations RWBD and RWBS at Time Window T=25

Since activity driven networks have power-law distribution as activity potential distribution, we ask in which ways it mirrors scale-free networks and in which ways it diverges from scale-free networks. Until now, we reviewed that scale free networks have almost power-law distributed degree, shortest-path betweenness and random walk betweenness distributions, and these curves for activity driven networks have also similar features with pareto distribution.

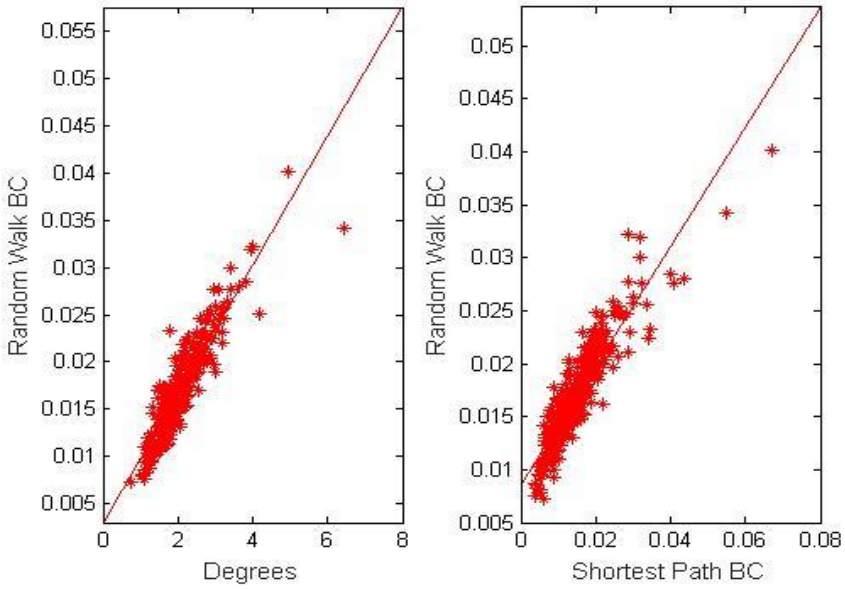


Figure 37: Plots of ADN for Correlations RWBD and RWBS at Time Window T=50

However, in figures 36 to 39 we notice that correlation behavior of ADN alters for the time windows that we observed. If we examine figure 36 the plot on the left seems more dispersed around the reference line than the one on the right. This means that the correlation of random walk betweenness with degree is significantly lower than the correlation of random walk betweenness with shortest path betweenness at T=25. These correlations are obtained for average of 20 networks and the same parameters as the section above.

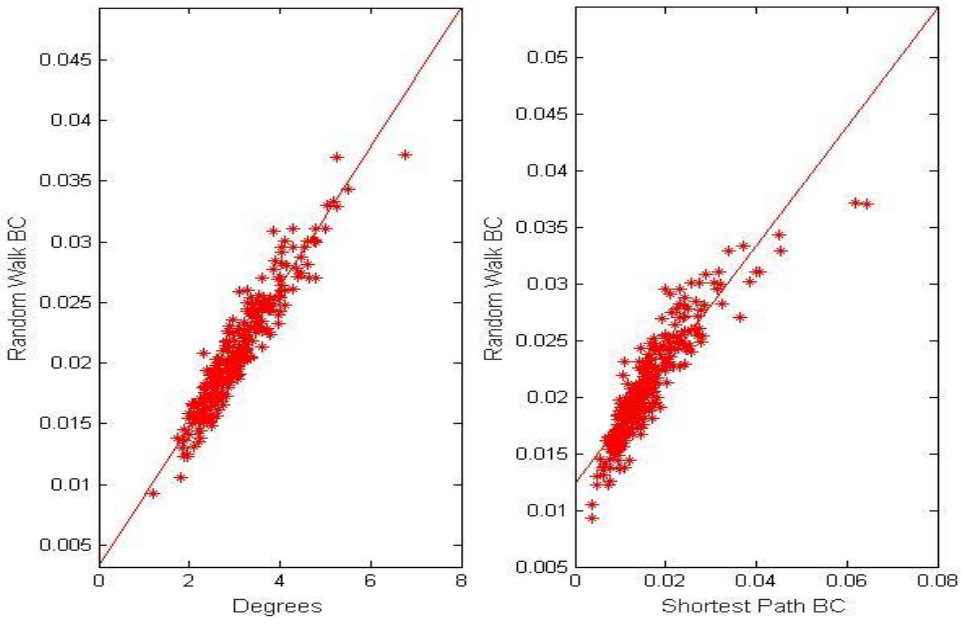


Figure 38: Plots of ADN for Correlations RWBD and RWBS at Time Window T=75

In figure 37 the points on plots on the right and left appear to be in similar distance from the reference line. This demonstrates that RWBD and RWBS are remarkably close to each other.

Finally, plots on the left side of figure 38 and 39 seem more orderly than the other. The correlation values related with the scatter plots are given with tableau 3.

Table 3: RWBD and RWBS Correlation Values for ADN at Different Time Windows

	RWBD	**RWBS**
T=25	0.7527	0.9623
T=50	0.9075	0.9065
T=75	0.9452	0.9042
T=100	0.9347	0.9094

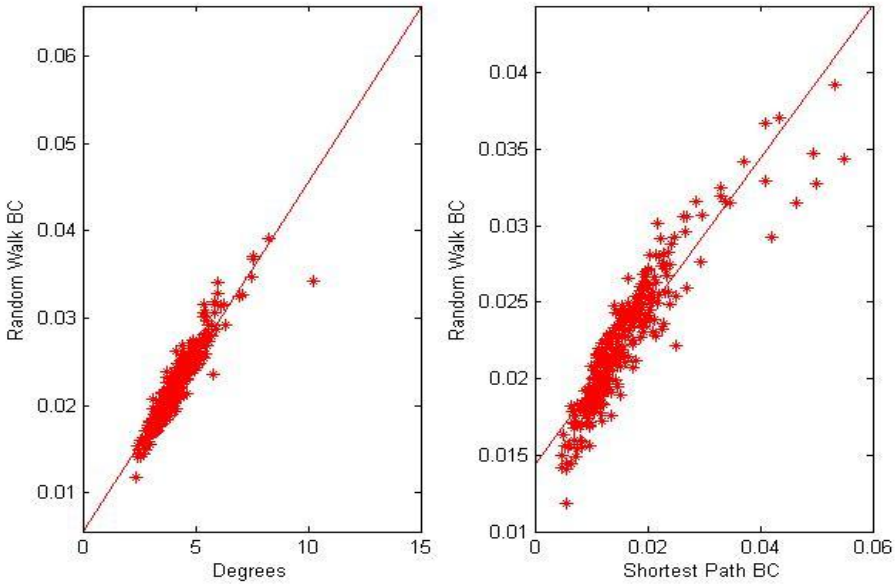


Figure 39: Plots of ADN for Correlations RWBD and RWBS at time window T=100

As a result, for small time lengths RWBS is greater than RWBD for activity driven networks, and as the time increases RWBS do not chance much while RWBD raises strikingly, almost 20 percent. At T=25 there are many disconnected groups in ADN, some nodes with high degrees are not connected through paths yet as we also learned from average path length plots. Therefore, RWBD is quite low. Later, as the separate groups become connected they cause an increase in RWBD. Meanwhile, RWBS is decreasing; this does not exactly mean that the network becomes less clustered since we discovered in section 2.2 that the clustering coefficient of ADN increases for greater T values.

In the next chapter we will discuss how this affects triadic closures in ADN.

4.4. Discussions

Activity driven networks have skewed degree, random-walk and shortest-path betweenness curves. Tales of these curves are similar to that of power-laws. Nevertheless, when we tested we found that they are not power-laws, since the curves are not completely linear n log-log plots. Moreover, we used Input Analyzer to test whether degree distributions of ADN at different time windows fit to any of random distributions. However, the test did not resulted with any good fit.

Correlation outcomes were also pretty interesting. RWBD correlation is low at T=25, and it increases for other time windows. The reason for that is at t=25 ADN is barely connected; there are a few disconnected groups. These separate groups might have nodes with high degrees, but since they are disconnected their random-walk centrality is quite low. These groups form links among each other later. On the other hand, the situation is vice versa for RWBS, it decreases for lower time lengths. This displays that ADN has small tight-knitted disconnected groups at T=25, and as the groups connect with links they also create new shortest-paths which are shorter from original paths.

Chapter 5

Effect of Multiple Mutual Friends on Link Formation for Activity Driven Networks

Social networks change over time by adding new edges, removing some old edges or altering the strength of ties. These changes are affected from many factors such as shared activities, similarities in individual's characteristics and shared friends. Triadic closure is one way to add new ties to networks and it is one of the basic principles of social networks. It means that if two people in a social network have a mutual friend, then there is an increased likelihood that they will become friends themselves at some point in the future(Easley and Kleinberg 2010)[8] for triadic closures. Thus, we can ask the question how the probability of triadic closures changes over time in evolving social networks. An advanced version of this question is what happens if two people have more than one friend in common? Does the probability of forming an edge between these two people increase, decrease or will it not be affected from change in number of mutual friends? Common sense would state that this probability will increase as the number of mutual friends raise. Then the next question comes, how does this

probability increases? In this chapter we will answer these questions for activity driven networks.

Kossinets and Watts (2006) [17] performed this analysis on a network which is produced from email communication dataset of a large US university. They observed the system for 355 days and gave link to two people if they exchange emails in 60 days. Thus, their integrated network always included 60 days in the past. The steps they used and we repeated for measuring the impact of multiple mutual friends for not connected two people are as follows:

1. First, we take a snapshot of the integrated network at a predetermined time window
2. Next, we calculate number of all not connected pairs who has exactly k mutual friends
3. We run the model for τ time steps and take a second snapshot.
4. Then we find the fraction of earlier identified pairs that are connected, T(k), and we plot k versus T(k).

Here, T(0) is the rate at which two people form a link who do not have any common friends and T(1) is the rate that triadic closures happens.

Kossinets and Watts(2006)[17] started their analysis from 61th day and they calculated T(k) each day until day 270 which means they chose 60 days as smoothing window. After the first calculation of T(k) at t=60, they took a snapshot of the network each day, and in each snapshot they looked back 60 days to identify number of non-connected pairs with k mutual friends. Next, they took the average of 209 T(k) curves and composed the plot in figure 40.

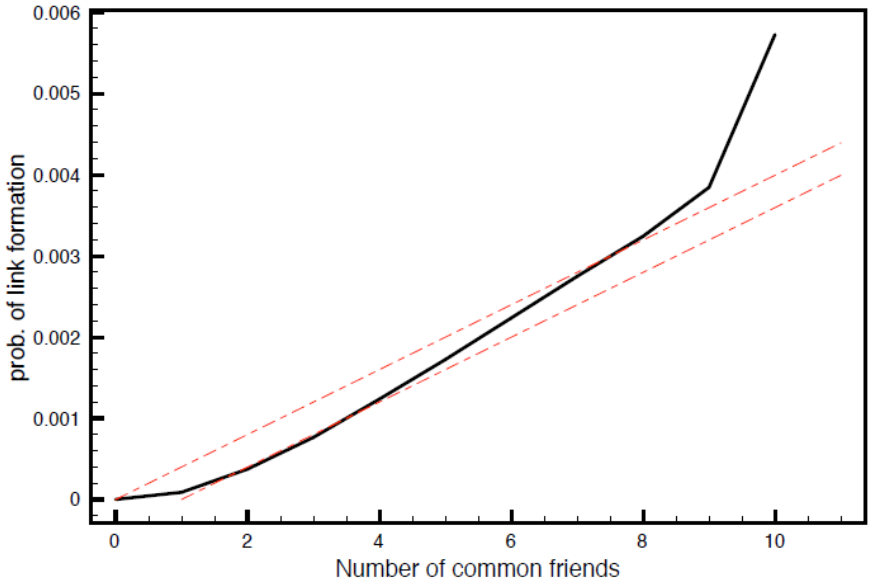


Figure 40: T(k)-k Figure for Email Communication Network in a US University[8]

If we look into figure 40 more carefully, the first thing we notice is that T(0) is almost 0, which means people who do not have any common friends barely form a tie. Furthermore, as the number of common friends increases so does the probability of forming a link. This is a proof for the existence of triadic closures. The shape of the curve is close to linear; however there are two distinguishing points, the first one is where the curve moves from 1 friend to 2 and the second one is at more than 10 common friends. The first turn states that having 2 friends in common impacts tie formation than 2 times more in comparison with having one mutual friend. Similar curve turns has been notified also in other datasets. (Backstromet al. 2006 [2], Leskovec et al. (2007) [18]). On the other hand, the second turn at 10 common friends is less significant since the data considerably shrinks there.

Upper red line corresponds to a baseline model, which is created to understand the plot better and we will call it Tbaseline. In order to explain this curve let us assume that the probability of forming a tie for two people depends on only their common friends. Each mutual friend that they have gives them an independent probability p of connecting a tie at each time step. In this case if these two people only have one mutual friend the probability of not forming a link is$(1 - \mathrm{p})$, if they have k mutual friends the probability of not forming a link is $(1 - \mathrm{p})^{\mathrm{k}}$, because of the independence assumption. Thus, the probability of them to form a link eventually is $1 - (1 - \mathrm{p})^{\mathrm{k}}$. As a result, we can draw the baseline curve as follows.

$$\mathrm{Tbaseline} = 1 - (1 - \mathrm{p})^{\mathrm{k}} \tag{22}$$

Easley and Kleinberg drew the second red dotted curve which is $1 - (1 - \mathrm{p})^{\mathrm{k}-1}$ with regard to show the effect of first common friend.

The baseline curve, which represents the independence situation, is almost linear, but it turns slightly downwards as k increases. Despite of their similarity, the real curve behaves in opposite way and turns marginally upwards, which validates that each extra common friend do not have independent effect on forming a tie.

5.1. Results for Activity Driven Networks

In this section we analyze the effect of number of mutual friends on triadic closures for activity driven networks. Generation of activity driven networks are completely markovian. Formation of new ties does not depend on existing links at any step. Looking from this perspective we expect T(k)-k curve of ADN to be similar to baseline curve.

There are several differences between our analysis and the research of Kossinets and Watts (2006) [17]. First of all, we do not limit the integrated network to number of days in the past. In other words, we do not remove any ties in the integrated network; we assume that once two people built a friendship they stay as friends through the entire study. Moreover, since at t=0 ADN consists of unconnected nodes, we set a smoothing window θ, and we start performing the calculations at $t = \theta + 1$.

Another attribute that we need to decide is sampling period, the length of time that we run the model after calculating number of non-connected pairs with k mutual friends and before calculation how many of them are connected. Kossinets and Watts (2006) preferred $\delta = 1$ day as sampling period, and they took the average T(k) for 60 days in order to plot figure 38. Furthermore for each smoothing window setting we measured T(k) for average of 50 different networks and 10 times consecutively.

Additionally, we select the ADN parameters N=300, m=2, η=10 and $\epsilon \leq x \leq 1$ with $\epsilon = 10^{-3}$ because of the reasons that we explained in earlier chapters.

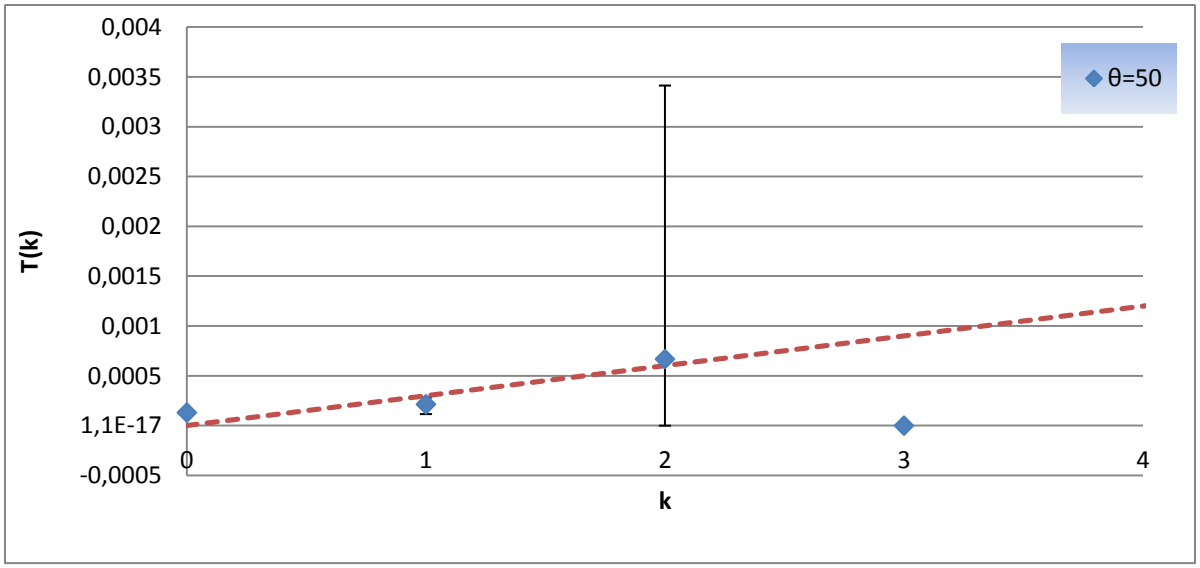


Figure 41: T(k)-k for ADN with Smoothing Window θ=50 and Sampling Period $\delta = 1$

In figures 41, 42, 43 and 44 given with solid lines plots of number of common friends(k)-probability of link formation(T(k)) for 1 time window sampling period and different smoothing windows, and the dotted lines are fitted Tbaseline curves. More specifically, in order to draw the curve in figure 41 first we run activity driven network model for θ=50 time steps and we count number of non-connected pairs for each k. Next, at T=51 found the fraction of those pairs who formed a link and calculated number of non-connected pairs again. Then, we found the fraction of connected pairs at T=52 again, and we continue this analysis until T=60. Figure 41 is constructed as the average of 50 different activity driven network and 10 different time steps from T=50 to T=60.

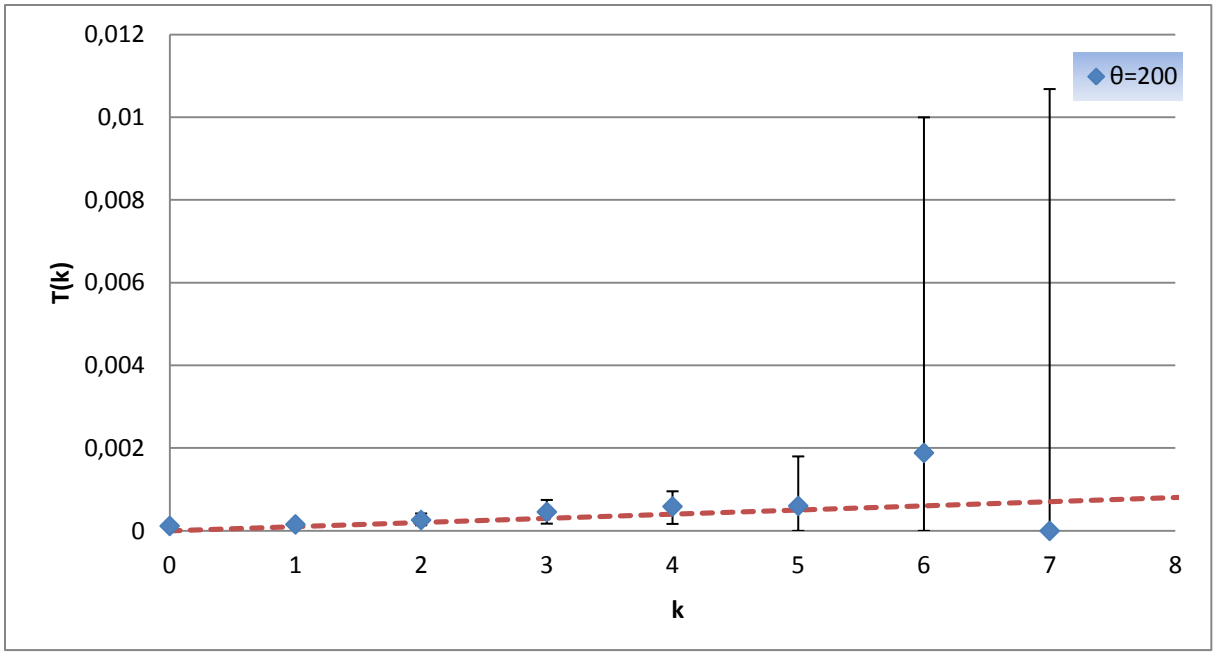


Figure 42: T(k)-k for ADN with Smoothing Window θ=200 and Sampling Period δ = 1

It requires great effort to eliminate all the noise in the data. Nevertheless, by taking average of 50 networks and 10 time steps we aimed to minimize effect of randomness.

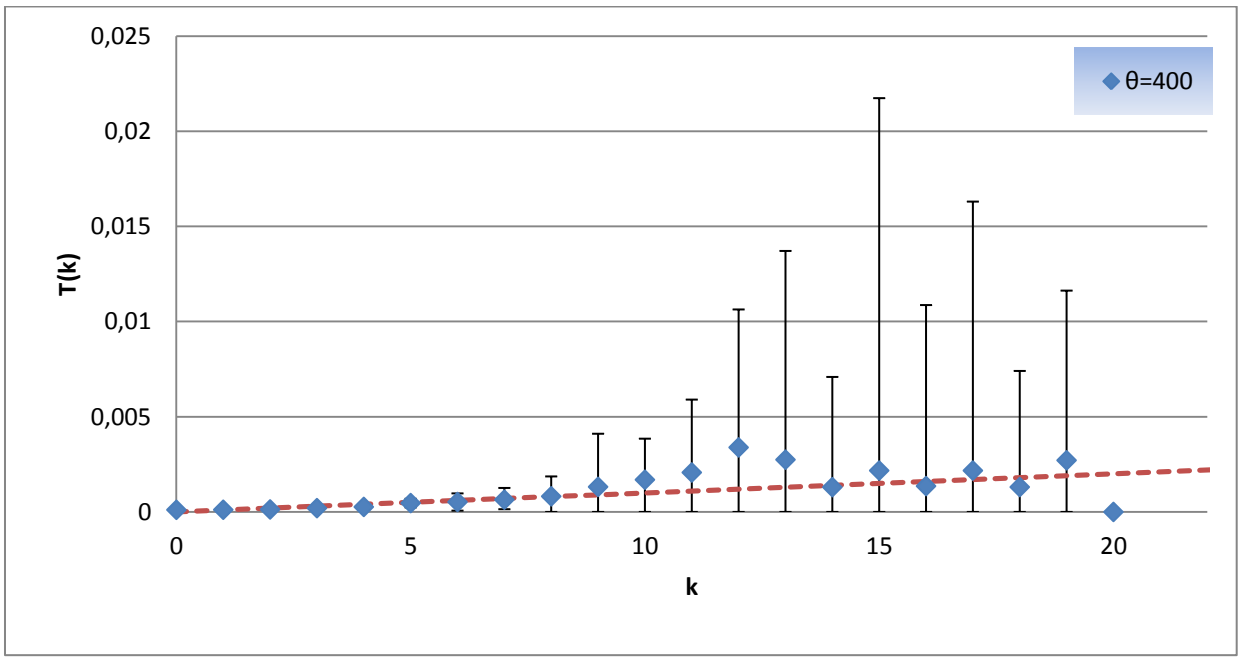


Figure 43: T(k)-k for ADN with Smoothing Window θ=400 and Sampling Period δ = 1

Looking closer to figures 41 to 44, and ignoring the noise in the data we notice that the curves have a close to linear shape which increases until a peak k value, and then they saturate and have more flat shape after that point. This shape is especially seen in figure 43, when we start calculating T(k) after θ=400. The probability of forming a link increases with number of common friends up to some point and then it does not have a significant change until it hits zero.

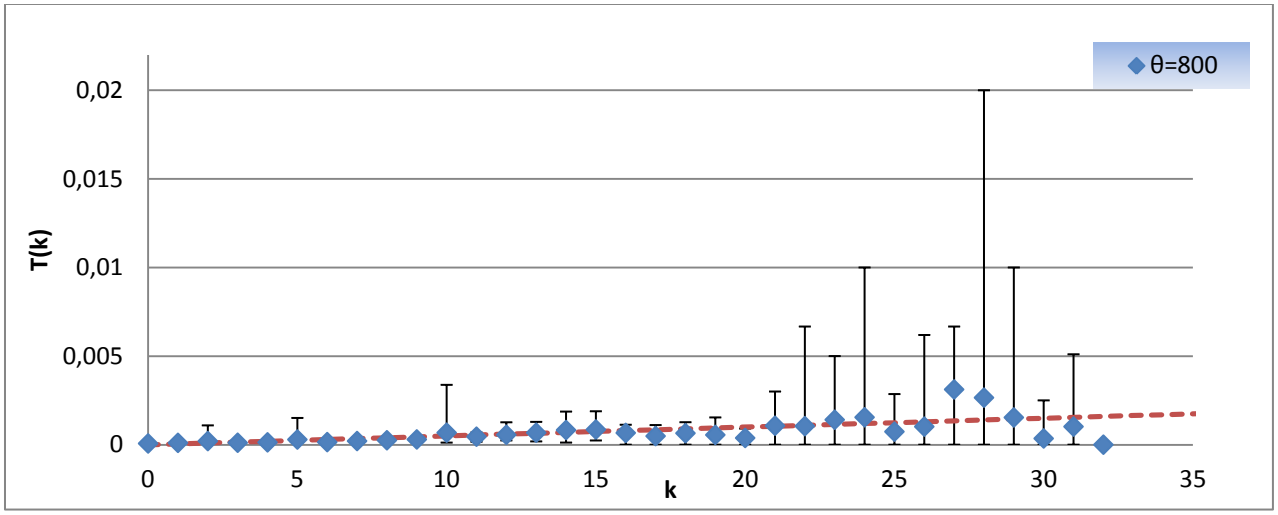


Figure 44: T(k)-k for ADN with Smoothing Window θ=800 and Sampling Period δ = 1

If we cut the curves from their peak points they resemble figure 40, they go upwards after following linear increase for a while, which means for higher values an increase in k cause a greater increase in probability of forming a tie, in comparison to lower k values.

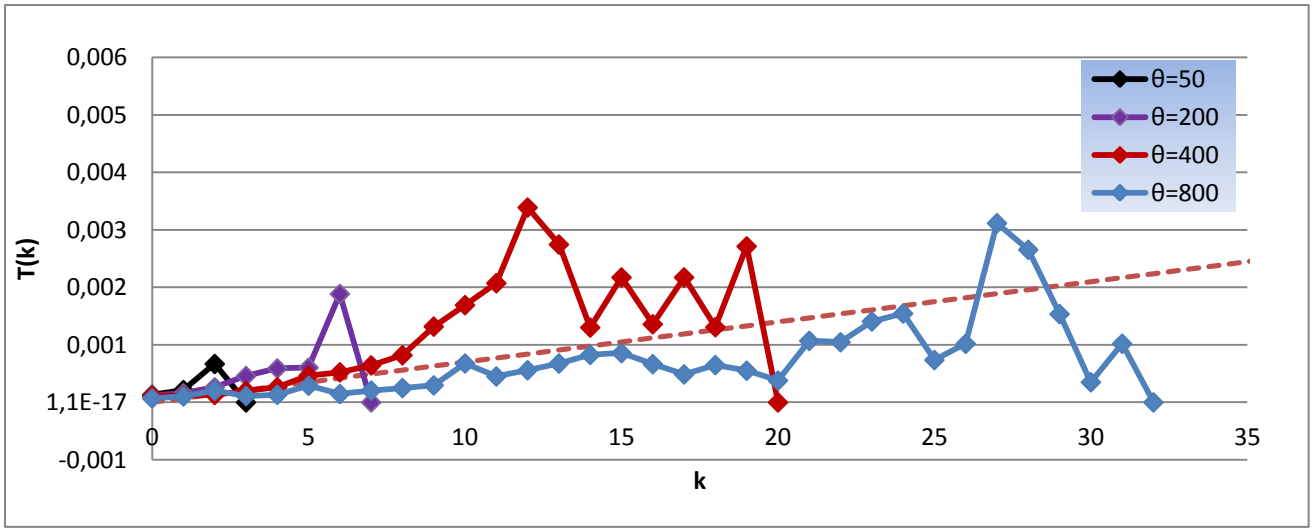


Figure 45: T(k)-k for ADN with Different Smoothing Windows and Sampling Period $\delta = 1$

In figure 45 we observe curves with different smoothing windows all together. Here, we recognize more clearly that as the smoothing period increases the slop of the T(k)-k curve decreases. Moreover, turning point moves forward and the curve stretches. Let us remember the discussion in section 4.3 about correlations among betweenness measures; RWBS correlation answers if most of the paths in a network are also shortest paths and if they are, this means for the network that it is highly clustered. When RWBS decreases, we expect that the probability of triadic closures declines too. Therefore, the fact that as length of smoothing window develops the slope of T(k)-k curve decreases, is because RWBS is getting smaller as the time window becomes greater.

One common feature as we also discussed for figures 41 to 45 is diminishing return effect [8]. Diminishing return effect represents the decline in the slop of curve after going upwards until a turning point k^*. It is detected for studies about membership closure, which means that a person who is part of an affiliation influences her friend to join the same affiliation; hence the second person eventually becomes a member of that society through her friend. The study of Backstorm [2] in which they measure the probability of joining a LiveJournal community as a function of number of friends who are already a members and the study of Crandall et. Al(2008)[5] in which they quantify similar analysis for editing Wikipedia articles .

Turning point k^* increases as we extend the length of the smoothing window θ. While in figure 41 for θ=50 k^* is 2, for θ=400 in figure 44 it is 12; thus as the smoothing window increases the curve shifts to right.

We included error bars for figures 41 to 44, for the purpose of observing the noise in the data more clearly. Error bars show the minimum and the maximum observed data for each point in the plots. For each parameter set since the number of data points declines as k grows, the error bars elongate for bigger values of k. However, we did not include them to figure 45, in order to see the difference in slopes of the curves more clearly.

All curves above in this section are not linear as Tbaseline, and T(k)-k analysis do not show independence for ADN. The fact that, at $t = t_1$ an activity driven network forgets all the edges that is formed at $t = t_0$, shows that for two non-connected people the effect of a common friend in forming a link is not independent from the effect of other common friends. For real life data, we can explain this situation with explanations from social life such as people have limited time to spend with their friends, so after a point an increase in number of mutual friends does not raises his motivation to form a connection. However, we are not dealing with a real life social network; our nodes do not have limited time to spend with their friends. Hence, the answer should be somewhere else.

We investigate the most important parameter of ADN to find an answer to this question, activity potential function. Kossinets and Watts [17] selected power-law distribution, since this way degree distribution of ADN resemble remarkably the degree distributions of three datasets that they presented. When we model ADN with power law distribution, nodes with highest activity potential will become active very often and eventually they will connect to everyone else. But the algorithm will continue to choosing nodes with high activity potentials, which means after some time length the algorithm will almost stop adding new edges to the network. Indeed, this is why we have a pick k point in T(k)-k curve, and diminishing return effect afterwards. Hence in order to quantify that we performed our analyses also on an ADN with uniform distribution as activity potential distribution. For this ADN most of the nodes have an average activity potential value. Therefore the algorithm will continue adding new edges to network until it becomes almost fully connected.

Below in figure 46 are the curves of T(k)-k for ADN with uniform distributed activity potential function for smoothing windows $\theta = 50$ and $\theta = 200$, and sampling period $\delta = 1$.

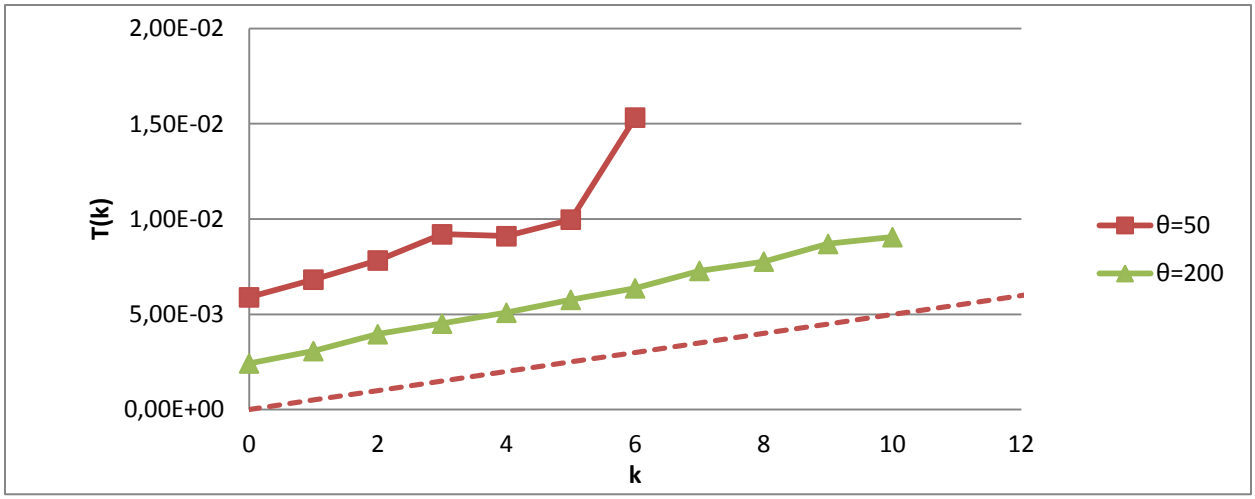


Figure 46: T(k)-k Curves for Uniform ADN with Smoothing Window θ=50 and θ=200, and Sampling period $\delta = 1$

Here we see a significantly differently scattered T(k)-k curve. There is no diminishing return effect or turning point k^*. In fact there is almost no change in the slope of curves through the entire scale. Their slopes are highly close to each other and both of the curves are parallel to Tbaseline as we see in figure 44 as if they are $1-(1-p)^{k-a}$ for some positive real number a.

5.2. Discussion

From chapter 5 we learned that activity potential function not only forms ADN's nature, but it also determines the impact of multiple common friends on link creation. Only if we model ADN with power-law distribution, regardless from its markovian structure, connections among nodes are dependent on number of mutual friends. As the number of mutual friends increase, the probability of link formation first increases and then reaches a saturation point. Therefore, if we benefit power-law distribution during ADN generation, the network immensely resembles real life social networks more than scale-free networks, and this makes activity driven networks quite eligible for performing social networks analyses without spending time on data analysis.

On the other hand, for uniform ADN we see an almost linear, increasing curve for both of smoothing windows, which means that each mutual friend has increase the probability of link formation with an independent probability p. ADN with uniform distribution activity potential

setting is more like random graph with an average degree 0.4 at a time step. Uniform ADN not only secluded from diminishing return affect, but it does also not have slope changes among curves of different smoothing windows.

Chapter 6

Conclusion

First part of this thesis is about shortest path betweenness and random-walk betweenness measures, what do they mean, how and where are they used, and how are they correlated to each other. We analyzed betweenness measures and degree centrality for different type of networks. Next, we detected the similarity in the curves of these centrality measures for the graphs we analyzed. Therefore, we also applied correlation studies among centrality measures. Our random-walk betweenness implementation of random graphs and scale-free networks displayed that; random-walk centrality is highly correlated with degree distribution for scale-free networks and less correlated with shortest-path betweenness, vice versa for random graphs. Hence, a node with a high degree in scale-free networks, also a node which is in between random walks very often.

Second and the main part of the thesis is about activity driven networks. Setup for generation of this model is assigning each agent an activity rate, which determines if this agent will form any links, and if so, on average how many links it ties in a time window. It fixes number of nodes from the beginning and only adds links through active nodes. We can estimate number of active nodes, number of links and average degree per iteration mathematically.

We inspect ADN at different time windows using centrality indicators and clustering coefficient. We discover that clustering coefficient of ADN is close to clustering coefficient

of scale-free networks, and significantly lower than clustering coefficient of random graphs. Degree distribution, random-walk betweenness and shortest-path betweenness of ADN have similar shaped curves; they are vast and left skewed in logarithmic scale.

We also examine correlations among centrality measures for ADN, and observe that especially correlation of random-walk betweenness and degree centrality changes remarkable with time window. We perform our analysis for four time windows and we explore that in this range as the time passes random-walk betweenness become more and more correlated with degree as in scale-free networks, but it becomes less correlated with shortest-path betweenness

We analyze centrality indicators with the purpose of understanding the triadic closures in ADN. We realize for ADN that if two people who are not connected with a link are more likely to form a tie if they have more common friends.

All the analysis we performed on activity driven networks indicates that connectivity pattern of ADN has strongly influenced by some real life social networks. Therefore, one can benefit it as a ground for solving network problems.

In activity driven networks nodes do not remember their earlier contacts, the model grows in a way that active nodes connect with uniformly selected other nodes to connect with at each iteration. This keeps the model simple and effective but maybe in the future studies one can include some memory factors to the equation.

Bibliography

[1] Albert, Réka, and Albert-László Barabási. "Statistical mechanics of complex networks." *Reviews of modern physics* 74.1 (2002): 47.

[2] Backstrom, Lars, et al. "Group formation in large social networks: membership, growth, and evolution." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

[3] Barabási, Albert-László, and Eric Bonabeau. "Scale-free networks." *Scientific American* 288.5 (2003): 50-59.

[4] Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." *science* 286.5439 (1999): 509-512.

[5] Crandall, David, et al. "Feedback effects between similarity and social influence in online communities." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[6] Diestel, Reinhard. "Graph theory. 2005." *Grad. Texts in Math* (2005).

[7] Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. science, 301(5634), 827-829.

[8] Easley, David, and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[9] ERDdS, P., and A. R&WI. "On random graphs I." *Publ. Math. Debrecen* 6 (1959): 290-297.

[10] Erd6s, Paul, and A. Rényi. "On the evolution of random graphs." Publ. Math. Inst. Hungar. Acad. Sci 5 (1960): 17-61.

[11] Freeman, Linton C. "A set of measures of centrality based on betweenness." *Sociometry* (1977): 35-41.

[12] Hage, Per, and Frank Harary. *Structural models in anthropology*. Cambridge University Press, 1983.

[13] Holland, Paul W., and Samuel Leinhardt. "Transitivity in structural models of small groups." *Comparative Group Studies* (1971).

[14] Holme, Petter, and Jari Saramäki. "Temporal networks." *Physics reports* 519.3 (2012): 97-125.

[15] Kapferer, Bruce, and James Clyde Mitchell. "Social networks in urban situations." (1969): 181.

[16] Karsai, Márton, Nicola Perra, and Alessandro Vespignani. "Time varying networks and the weakness of strong ties." *Scientific reports* 4 (2014).

[17] Kossinets, Gueorgi, and Duncan J. Watts. "Empirical analysis of an evolving social network." *Science* 311.5757 (2006): 88-90.

[18] Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman. "The dynamics of viral marketing." *ACM Transactions on the Web (TWEB)* 1.1 (2007): 5.

[19] Newman, Mark. *Networks: an introduction*. Oxford University Press, 2010

[20] Newman, Mark EJ. "A measure of betweenness centrality based on random walks." *Social networks* 27.1 (2005): 39-54.

.

[21] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69.2 (2004): 026113.

[22] Newman, Mark EJ. "The mathematics of networks." *The new palgrave encyclopedia of economics* 2.2008 (2008): 1-12.

[23] Newman, Mark EJ. "Clustering and preferential attachment in growing networks." *Physical Review E* 64.2 (2001): 025102.

[24] Perra, Nicola, et al. "Activity driven modeling of time varying networks." *Scientific reports* 2 (2012).

[25] Potterat, J. J., et al. "Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs." Sexually transmitted infections 78.suppl 1 (2002): i159-i163.

[26] Vespignani, Alessandro. "Modelling dynamical processes in complex socio-technical systems." *Nature Physics* 8.1 (2012): 32-39.

[27] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world' networks." *nature* 393.6684 (1998): 440-442.

[28] Zachary, Wayne W. "An information flow model for conflict and fission in small groups." *Journal of anthropological research* (1977): 452-473.