Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar

Geometric explanation of rich-club behaviour in complex networks

Author
Csigi Máté

Supervisors
dr. Gulyás András, Kőrösi Attila, dr. Bíró József

October 23, 2016

Abstract

There are a wide range of models capable of generating scale-free networks with a given degree distribution. Another important parameter of networks besides the γ parameter describing the degree distribution is the rich-club coefficient. However, there are no models which could modify this parameter of graphs in an intuitive way. In the present paper we propose a network generating model which can generate random networks with a given rich-club coefficient.

Chapter 1

Introduction

Networks of all sorts are around us including the Internet, power grid networks, transport networks and even our very bodies. The interactions of our cells and proteins can be understood from a network science perspective. What is interesting is that all these systems are very similar to each other when regarded as networks. One of the most striking similarities they all share is that their degree distribution is scale-free with exponent (γ) between 2 and 3.

There are many models capable of generating random graphs which have scale-free degree distribution with adjustable γ parameter. These models are capable of generating real-network-like graphs. One of the most famous of these models is the Barabási model. There are others ones using geometric aspects to generate such graphs, since (hyperbolic) geometry seems to describe some properties very deeply.

Another, yet not so wide-spread notion is that of rich-club coefficient. Contrary to other metrics (e.g. degree distribution, clustering, diameter), in which real networks exhibit surprising similarity, the rich-club organization is something that makes networks look way different. In short, the rich-club coefficient how tightly the hubs (nodes with many links) are connected to each other. If the rich-club coefficient is big means that there is some sort of an oligarchy present in the network. The notion applied to a social network would mean that those who have many friends also know each other well. However other networks (e.g. the power grid or, protein networks) seems to lack such rich-clubs.

Despite the high volume of available network models, one couldn't find a single one generating graphs with modifiable rich-club coefficient in a simple, intuitive way. This would be very helpful though, because with a model like that we could fine-tune the networks not only varying the degree distribution but the rich-club coefficient as well. In this work we present a model with the above mentioned capabilities.

Chapter 2

Overview of network theory

The set of nodes is the fundamental unit of networks. These nodes are connected among each other with edges which mean some sort of a connection between the nodes. In case we talk about a social network the edge can mean friendship between two people, but if the nodes are molecules then an edge means a possible chemical interaction between them. The edges can be directed or undirected and we can also assign weights to them signifying the strength of the interaction between the two vertices. In this paper we will mainly discuss undirected networks.

In the following I will present the most important features that we can use to analyze and characterize real-world networks. There are quite a few such notions which explore graphs from a different point of view each one offering a different perspective on the network in question. There are many ways to look at networks each one revealing other properties other measures may not cover. That's why it's important to analyze different aspects of real graphs. The most influential ones are defined below.

2.1 Properties of real networks

The degree of a node is the number of nodes which are connected to it. It simply measures how well-connected a given node is. Applied to a social network a node with low degree would mean someone with hardly any friends. In case of directed networks we make difference between in-degree and out-degree based on the distinction of edge direction. In undirected network we simply count the edges associated with a node. In many cases it is important to have a general view of the degrees of nodes present in the network. Keeping count of all nodes with their degrees in a table-like manner is hard to work with. That's why some useful measures are introduced to make it easier to obtain a general view of the network's degree properties.

2.1.1 Degree distribution

We define n_k as the number of nodes with degree above k. The relation between n_k and k is the degree distribution of the given network, one of such useful notions revealing some crucial properties of graphs in an informative way. In almost all real-life networks the degree distribution is scale-free meaning that the n_k-k relation is a power law ($n_k \sim k^{-\gamma}$) function. Such plots can be seen in Figure 2.1 where the degree distribution of some networks have been plotted. The degree distribution is very informative, because we can understand some basic properties of networks at a glance. It can be observed that all degree distributions are approximately lines in the log-log plot, hence follow power law functions. The exponent of the power-law function (γ) of the plots are varying between 2 and 3. A scale-free distribution intuitively means that most of the nodes have a low degree, but it's not uncommon to have some with outstanding degrees. If the degree distribution would be exponential it would mean that there is almost no chance to have nodes with significant number of connections.

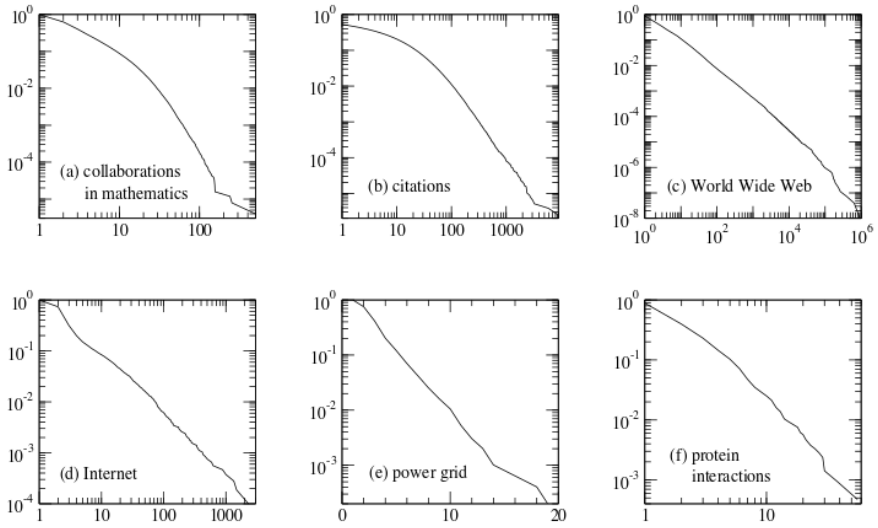

Figure 2.1: The degree distribution of some real networks are plotted. All have a power-law degree distribution. The figure is taken from Newman's monography on networks [1].

2.1.2 Diameter

The diameter [1] of a network is defined as the length of the longest of all shortest paths between the nodes. All networks observed around us have a relatively small diameter thus they have the so called small-world property meaning that from any node you can get to any other in a small number

(proportional to the logarithm of the number of nodes in the graph) of steps. This concept projected to a social network means that a random person chosen from Budapest is likely to know one of your friends of your friends. Karinthy estimated that the network of all people on Earth has a diameter of 6 meaning that if you would like to send a letter to someone in Japan you have never met by sending it to one of your friends who send it to one of his friends and so on then your letter can get to its destination in 6 steps. Experiments like the one described here have been carried out in the US confirming this surprising property. Having a small diameter is crucial is real networks as it allows information to travel fast between nodes. In Figure 2.2 the notion of small-world property is presented in a graphical manner showing the average distance in real-life graphs plotted against the size of the network.

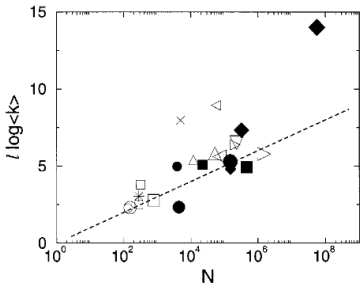

Figure 2.2: In this figure the average distance is plotted of some real networks to demonstrate that the average distance is quite small compared to the network size. The figure is taken from one of Barabási's articles [6].

2.1.3 Clustering coefficient

The next definition to be presented is clustering which is also called transitivity. The intuition to this comes from the observation that if node A is connected to B and B is connected to C then A is likely to be connected to C as well. So clustering measures the strength of groups formation. In the context of social networks it would mean that two of my friend are likely to be friends themselves. The formal definition goes like the following:

$$C = \frac{\text{3x number of triangles in the network}}{\text{number of connected triples of vertices}}$$

Where a connected triple means a single vertex with edges running to an unordered pair of others. In effect clustering measures the ratio of full triangles to the number of three-node formations with 2 or 3 edges amongst them. The interesting thing about this measure that when calculated for

Network	Clustering	Rand. graph clust.	Multiplier
Internet	0.24	0.00060	400
power grid	0.080	0.00054	148
mathematics collaborations	0.15	0.000015	10000
word co-occurrence	0.44	0.00015	2933
metabolic network	0.59	0.090	7

Table 2.1: In this table the clustering coefficient of some real networks is compared to the clustering of random graphs with the same number of nodes. The *Multiplier* column shows how many times the real network's clustering is greater than that of the random graph. The results are taken from one of Newman's articles [2].

social and other types of networks yields a considerably greater value then for a random network. This is further extended in Table 2.1 where real networks are compared to random graphs from this perspective. So clustering seems to reveal some very fundamental features of networks.

2.1.4 Rich-club organization

The final but crucial definition I will give is that concerning the rich-club parameter [3]. This parameter describes how the hubs of the network are connected to each other. A big value indicates the presence of a kind of oligarchy in the network, that is high-degree nodes are connected to each other. If the rich-club coefficient's value is small it means that influential elements in a network don't know each other. We define r_k as

$$r_k = \frac{\text{number of edges in } G_k}{\text{number of all possible edges in } G_k}$$

Where G_k represents the subgraph of G which only contains the nodes whose degree is greater than k. However these values are usually normed with the ones we would get from a random graph. With this slight modification it is easier to distinguish between the different kinds of rich-club coefficients. In contrast to the above mentioned measures of networks the rich-club coefficient is really different for different real networks. This huge diversity can be seen in Figure 2.3.

The primary motivation of graph generating models is to give a way to make networks which are very similar to real ones. The network's measures defined in this chapter all serve the function of comparison that can tell how close are real networks to our generated graph. There are a number of models for generating networks each of which explain a different set of

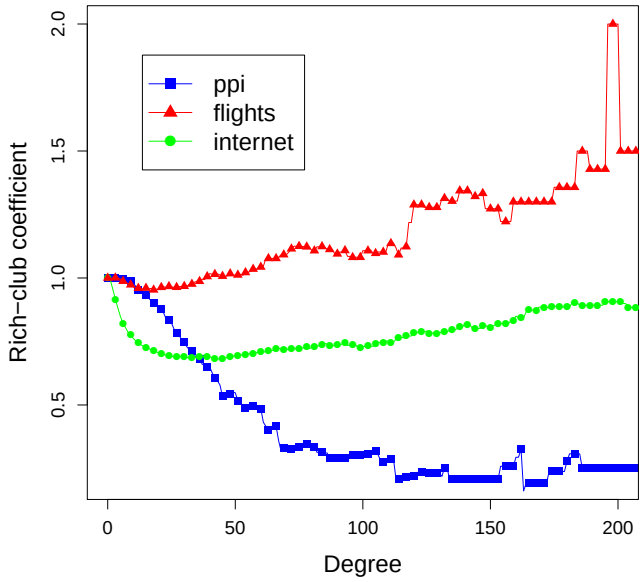

Figure 2.3: This figure shows rich-club coefficient plots from real world graphs. The rich-club coefficients of these three graphs are really different while their degree distributions are highly similar.

properties of real-life networks. With the help of the we could understand some important processes of networks formation.

2.2 Network models

2.2.1 Erdős-Rényi model

The first model I will present is the Erős-Rényi model [4]. This model is using two parameters to generate random graphs: the number of nodes and the number of edges. The nodes are connected to each other by the edges each possible edge being equally likely until we reach the predefined number of edges. One such graph generation could go like the following: We have 6 points and we would like to make a random graph from them using 10 edges. Let's decide where the edges will go by throwing two dices. The numbers of the dices will tell the two nodes the edge will connect. In case the two numbers are the same or the edge already exists then we throw again, since we would like to avoid self-loops and double edges. So we start by throwing the dices. We get numbers 2 and 5, so we connect nodes 2 and 5 with an edge. We have 9 edges left to add. We keep doing this procedure until we have used all of our edges and then we have our Erdős-Rényi random graph.

It is interesting to see that such a simple model can itself account for some central properties of real networks. From the detailed analysis of this model the relatively small diameter of the resulting graphs come out but it fails to explain the formation of hubs and clustering observed in real-life networks.

2.2.2 The small-world model

The Watts and Strogatz model [5] was designed to address these two primary limitations of the Erdős-Rényi model. The algorithm goes like equally placing N nodes on a circle and connecting each one to its k closest neighbors ($k/2$ on each side). Then for each node take each of its edges one after another and replace it with probability β. The new edge's one vertex stays at that given node from which it was evaluated and the second one is placed to some other node with equal probability avoiding self-loops. So each time we get to a new edge we calculate a random number in the $[0, 1]$ interval. If our random number is smaller than β then we decide on the new target node by the throwing of an N-side dice the starting node staying the same. With β parameter converging to 1 the model reproduces the Erdős-Rényi model, because all edges are replaced randomly none of them staying at its original place. The formation of the graph during this process can be followed in Figure 2.4. The produced graph has a very small diameter scaling linearly with the system size. The clustering coefficient also approaches the one observed in nature explaining groups formation. The drawback of the system is that it produces an unrealistic degree distribution which we earlier saw was a crucial component of a nature-like graph.

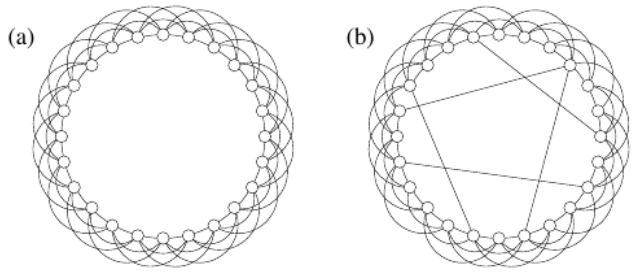

Figure 2.4: The Watts-Strogatz model in function at two different stages. (a) The nodes are placed on the circle and each one connected to its k closest neighbors. (b) Using the p parameter some edges are replaced.

2.2.3 Barabási-Albert (BA) model

What follows is the Barabási model [6] which generates scale-free graphs using a preferential attachment mechanism. This model incorporates two

concepts which exist widely in nature: growth and preferential attachment. Growth in this case means that we don't have all of our nodes at hand during the whole process, but nodes come one after another. The concept of preferential attachment is that the nodes already present in the network have a sort of wealth quantity and nodes are more likely to attach to those that are wealthier. In this case this wealth quantity is the degree of the given node. The algorithm begins from an initial network and adds nodes one after another and connects them to m other nodes in the network. The possibility that the new node will be connected to the node is proportional to its degree. Thus the nodes with higher degree have more chance to get new connections and that's how hubs are formed in the system. The Barabási model explains the small diameter in real networks and produces scale-free degree distribution.

2.2.4 Malkov's model

There is a geometric model which can reproduce these results proposed by Malkov [7]. Malkov's model goes like the following: to an Euclidean disk of radius R add points one after another to random locations equally distributed on the disk. Upon adding a new node it is connected to it's k closest neighbors. The distance is calculated by dividing the Euclidean distance by the square root of the degree of the target point. This distance definition is based on the intuition that nodes with big degrees are closer to other points in some sense. What results from this is a scale-free network with a γ parameter of 3. The k parameter's value is equal to the average degree of nodes in the network, because in each iteration k new edges are added.

2.2.5 Summary of network models

Table 2.2 summerizes the different properties of these networks showing which network properties they can explain. Network generating models is a rapidly developing brach in network science with new models appearing each few months explaining real networks in a different way. There are models which are grounded in hyperbolic geometry taking advantage of a different kind of space while other approaches use fractals with much success but we shall not dwell on such matters here. My model extends Malkov's approach with a small, yet remarkable step which allows changing the rich-club coefficient of the resulting network leaving all the other properties unchanged.

Network	Degree distr.	Small-world	Clustering	Rich-club
Erős-Rényi		X		
Watts-Strogatz		X	X	
Barabási	X	X		
Malkov	X	X	X	
Our model	X	X	X	X

Table 2.2: Comparison of network models.

Chapter 3

The Model

The basic idea is that long edges are infeasible so they are only implementable by adding also a bridge node to the midpoint of the edge. A practical example of this notion comes from the power grid network, where the power cannot be transmitted effectively to long distances without transformation.

The generation of the graph is determined by four parameters, namely the number of nodes (N), radius of the disk(R), the connection parameter(k) and the threshold parameter(th). The N parameter determines the number of nodes in the network. The total number of nodes will be N plus the bridge nodes we add to cut long edges in half. The R parameter is the radius of the disk on which the nodes will be equally distributed. The k parameter says the number of the closest neighbors to which a new node will be connected. Finally, the th parameter is the one which serves as a distance limit for new edges. If an edge is longer than that then it will be cut in half with a new node placed in the middle. This parameter was not originally present in Malkov's model and is at the core of our model.

The pseudo code of the model can be seen below.

Algorithm 1 The Model - parameters:N, R, k, th

```
for i=0, i<N do
    add a new node by placing it to a random location on the disk
    angle = random(0, 2π )
    radius = √(pR²) , where p =random(0, 1)
    for j in k closest neighbors do
        dist_j = EuclideanDistance(i, j)/√(degree_j)
        if dist_j < th then
            connect them directly
        else
            place a node to the midpoint
            connect both nodes to this extra node
```

In Figure 3.1 a network generated by this model is plotted. It can ob-

served in the figure that there are only a few nodes with outstanding degrees the majority of nodes having small connections.

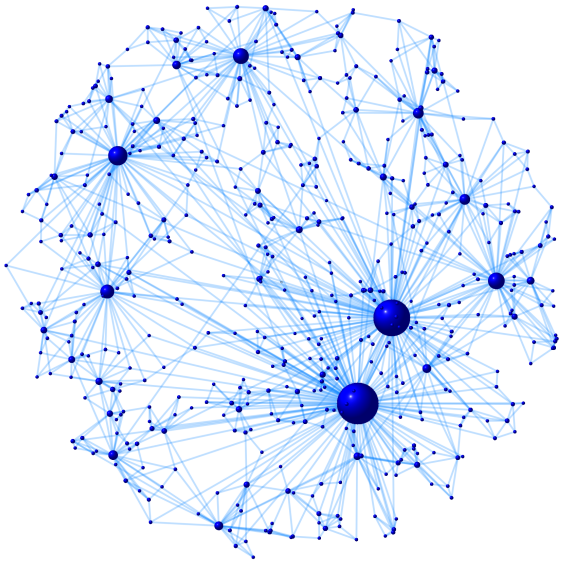

Figure 3.1: A network generated by the model. The size of the node is proportional to its degree.

The time evolution of the model is presented in Figure 3.2 where two added nodes are examined. It gives an intuitive explanation that no extra bridge nodes will be added after a certain number of iterations.

In case we set the threshold parameter to a large value we get back Malkov's original model because we will never insert new nodes to the graph. Our expectation is that these extra nodes will play the roles of bridge nodes in the system. The bridges nodes are those which connect hubs. In many cases in real-life networks great hubs are connected via one (or few) bridge nodes whose degree is considerably smaller than that of hubs. These special types of nodes have a crucial role in the dynamics of the network for this very reason connecting hubs. The threshold parameter is the tool in our hands with which we can influence the rich-club coefficient of the graph.

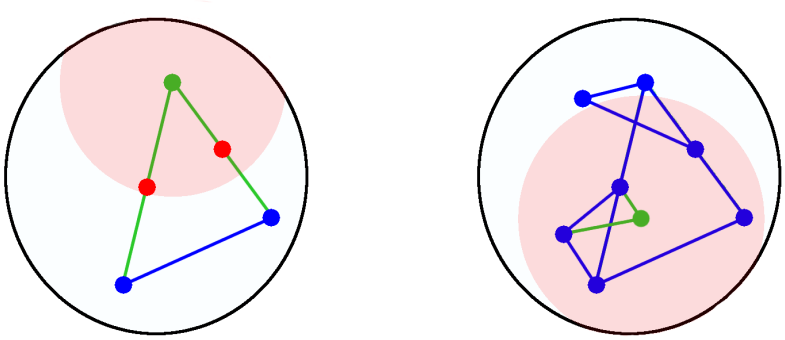

Figure 3.2: In the left figure initial nodes in blue are present in the network when we add a third one (green). The red circle represents the threshold around the new node. Either of the two blue nodes is within the threshold so extra nodes are (red) are inserted to the midpoint of the edges. While in the right figure which shows us the same network in a later stage no extra node is inserted, because there are many nodes falling within the threshold.

Chapter 4

Results

In this section we present a detailed analysis of the graphs obtained from the model. We will analyze three representative simulation results of threshold parameter of 6, 30 and 60 on a disk of radius 30. We will compare these networks to each other as well as to real networks to find out what the model is capable of.

4.1 Degree distribution

The obtained networks have a scale-free degree distribution. In Figure 4.1 the degree distributions of three networks which were generated with our model with different threshold parameters are plotted. It can be seen that the degree distributions don't differ remarkably from each other, because the threshold parameter in not modifying the degree distribution.

A way to prove that the degree distribution won't differ remarkably from that of the Malkov model is that the number of bridge nodes we added are very small in number compared to the number of nodes in the network. For this reason the overall degree distribution of the graph does not change considerably. The approximated number of bridge nodes at each iteration of the network generation can be calculated as follows.

Lemma 1. *The number of bridge nodes converges to a relatively small value compared to the network size during the generation of the graph.*

Proof. The following function gives us the probability that if we already have j nodes in the network and another node is given there will be $1, 2...N-1$ nodes within a circle of radius th.

$$\min(N-i, j-i)\binom{j}{i}p^i p^{j-i}$$

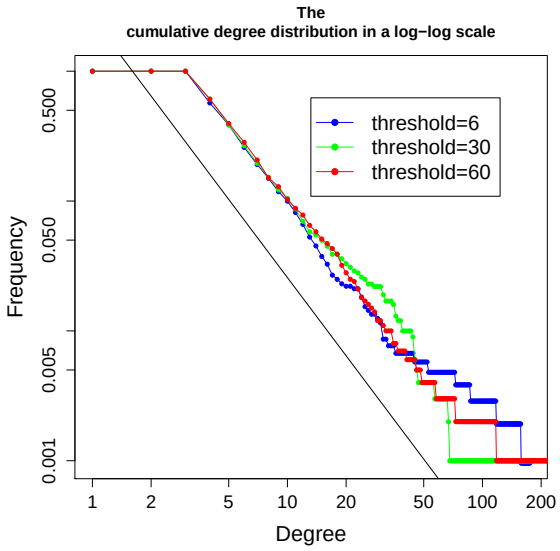

Figure 4.1: The degree distributions are plotted for different threshold parameters generated by the model. The degree distributions of these networks are almost the same.

Where N is the number of nodes we want to have in the graph, j is the number of nodes we already inserted, p is a probability which comes from the threshold parameter. The idea is that we assume that the nodes already placed on the disk are distributed equally. From this follows that when we add a new node the probability that another node is closer to it then the threshold value (th) is the area of a small circle of radius equal to the threshold parameter divided by the area of the full disk. This is how we calculate the p value. And finally the binomial distribution gives us the probability that a new bridge node needs to be inserted at all.

If we multiply it with $N - i$ as sum it based on i then we obtain the estimated number of bridge nodes at this stage of generation.

$$\sum_{i=0}^{\min(N-i,j-i)} \min(N-i, j-i)\binom{j}{i} p^i p^{j-i}$$

□

One more aspect to take into consideration is that the distance is calculated by dividing the Euclidean distance by the square root of the node. This means that the radius of the circle should not be the same during the whole process, but should grow proportional to the average degree of the

graph. Such growth would indicate that the nodes are closer to each other in a sense resulting from the degree-growth of nodes. This concept is taken into account in the calculation resulting in a varying value of p.

In Figure 4.2 we plotted the calculated theoretical number of bridge nodes in each iteration together with the simulation result for the same parameters. The two plots are really close to each other.

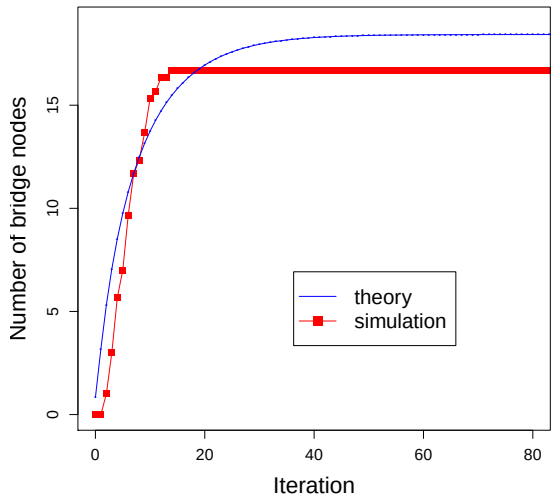

Figure 4.2: The theoretical number of bridge nodes in each iteration is plotted with the simulated values.

4.2 Clustering

The generated networks have a high clustering-coefficient with a value very close to real networks. When compared to each other the difference is so slight that we can say that the threshold parameter does not change the clustering coefficient of the networks having effect only on the rich-club parameter.

Network	Clustering coefficient
Generated threshold=6	0.57
Generated threshold=30	0.67
Generated threshold=60	0.68
Metabolic network	0.67
Physics Co-authorship	0.56
Internet	0.39

4.3 Diameter

The diameter of all three generated graphs is equal to 8, which is relatively small compared to the number of nodes being 500. This shows that the model produces networks with the small-world property reproducing one of the central features of real networks.

We also analyze the diameter and average distance in the resulting numbers for varying node numbers. The calculated values are plotted against the theoretical expectations. Both the diameter and the average distance should grow logarithmically with the number of nodes. The plot can be seen in Figure 4.3.

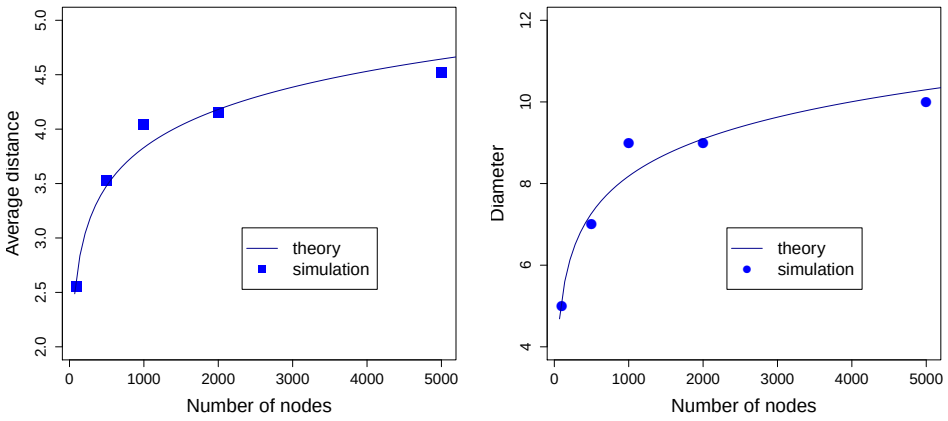

Figure 4.3: The average distance of the network is plotted against the number of nodes. The simulation results clearly show that a logarithmic function fits on them.

4.4 Rich-club coefficient

The simulation results in Figure 4.4 clearly show that the generated graphs differ greatly in their rich-club coefficient depending on the threshold parameter we set for it. We can generate graphs with a big diversity of rich-club coefficients mirroring real world networks which also show a great diversity in this respect. While the protein-protein interaction network has a really low rich-club coefficient meaning that the hubs are not well connected to each other, the flight graph has a significant rich-club value. But there are many examples for different rich-club values in nature. This diversity is demonstrated in the left of Figure 4.4 where there networks are examined from this point of view. On the right a very similar picture can be seen. These plots however were obtained from the graphs generated by our model, show-

ing how it can produce very different networks by adjusting the threshold parameter.

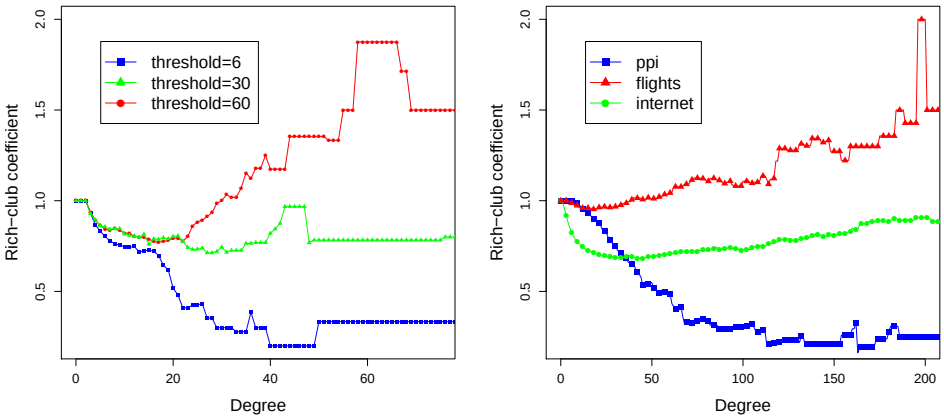

Figure 4.4: The left figure shows three plots of rich-club coefficient from graphs generated by the model with different threshold parameters. With the help of the model we can generate graphs with rich-club parameters in a wide spectrum imitating real world graphs. While in the right figure a very similar plot can be seen, but taken from real networks. This plot demonstrates that the model is capable of generating graphs with really different rich-club coefficients imitating this important property of real networks.

This is the main result of the paper presenting that apart from preserving all important features of real networks present in previous models the model is capable of modifying the rich-club parameter of the resulting graphs. Our geometric approach explains how it is possible to set the rich-club parameter of the graph by setting a single parameter.

Chapter 5

Conclusion and future work

There are many models which can generate random graphs that mimic real networks in one way or another. Yet no remarkable improvement has been made in the direction of generating graphs with adjustable rich-club coefficient. However, the rich-club coefficient seems to be an important aspect of network analysis, because there are great differences in rich-club coefficient values among real networks. So this parameter could possibly account for the diversity in networks we see around us.

In the paper we presented a new network model, the modification of a homophilic method, which can generate networks with a big diversity of rich-club coefficient values while preserving crucial properties like scale-free degree distribution, small-world property and high clustering present in real networks. Our model generates scale-free graphs with a γ parameter close to 3. We would like to improve on the model that would allow the γ parameter to be modified. The concept of cutting long edges in half by placing a node between can by applied to other models as well for example to some using hyperbolic geometry. Some early results show that it can be a promising direction as well.

Bibliography

[1] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[2] Mark EJ Newman. Random graphs as models of networks. *arXiv preprint cond-mat/0202208*, 2002.

[3] Vittoria Colizza, Alessandro Flammini, M Angeles Serrano, and Alessandro Vespignani. Detecting rich-club ordering in complex networks. *Nature physics*, 2(2):110–115, 2006.

[4] P Erdős and A Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[5] Duncan J Watts and Steven H Strogatz. Collective dynamics of âsmall-worldânetworks. *nature*, 393(6684):440–442, 1998.

[6] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[7] Yury A Malkov. Growing homophilic networks are natural optimal navigable small worlds. *arXiv preprint arXiv:1507.06529*, 2015.