Complex Network Analysis of Darknet Black Market Forum Structure

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

PHYSICS

Author :	Toms Reksna
Student ID :	s1760998
Supervisor :	Diego Garlaschelli
2^{nd} corrector :	Frank Takes

Leiden, The Netherlands, September 18, 2017

Complex Network Analysis of Darknet Black Market Forum Structure

Toms Reksna

Huygens-Kamerlingh Onnes Laboratory, Leiden University
P.O. Box 9500, 2300 RA Leiden, The Netherlands

September 18, 2017

Abstract

This work examines the network structure of illicit marketplaces that operate on the darknet. These on-line marketplaces are crawled to obtain data of inter-user communications and this data is parsed in a network structure and its physical properties are analysed. The Configuration Model is used as a null model to investigate the patterns in these networks to reveal information about their topology. This information is applied to interpret the behaviour of users within these illegal marketplaces.

Contents

Chapter 1

Introduction

In the recent years a whole new type of crime has emerged - international trade of illegal merchandise on online black marketplaces also known as cryptomarkets. Within these marketplaces users trade illegal goods including, but not limited to: drugs, weapons, child pornography, hitman services and others. The vendors and buyers protect their identity, masking their IP addresses by using the TOR network and purchasing the goods via the decentralized virtual cryptocurrency Bitcoin[1].

These two technologies have allowed to guarantee almost perfect anonymity and this has made cryptomarkets the main choice for illegal trade activities for sophisticated criminals.

1.1 Darknet black markets

The rise of these illicit sites was pioneered by a darknet site called ”Silkroad”, which was operational from February 2011 to October 2013, when it was shut down by the US Federal Bureau of Investigation (FBI) [4]. This attracted a huge media attention to this case, which lead to an increase in popularity and this method of exchange of illicit goods became known to wider range of public. At the time of writing this thesis there are more than 20 darknet black markets that are operational.

1.2 Black market community as a network

The research done on drug distribution networks has shown that the old criminal structures dominated by pyramid-shaped bureaucracies nowdays are relatively rare and in

turn decentralized groups are operating the drug trade. As conditions change dynamically, individuals whithin these groups form and dissolve relationships based on emerging risks and oppertunities. With new oppertunities of moving their business online, these communities have evolved to have more direct connections between consumers and drug producers decreasing the number of intermediary nodes and thus increasing network efficiency [10]. This is why it is important to study these online black marketplaces, in search for any additional information that might potentially help combat an increasingly efficient and very large criminal network.

1.3 Physical analysis of networks

In network science quite often models of random graph generation are constructed in order to understand better the dynamics behind the real world networks. One such model studied in depth in this work with respect to the black market community networks is the Configuration Model. This work tries to see whether or not this model is sufficient in explaining how communication propagates throughout these communities. In essence the main research question is whether given the amount of participants in these networks and the amount of communication in-between them, is everything else random, or is there additional structure in the network? (This question is more concretely defined in the coming chapters)

Chapter 2

Data extraction

In my original darknet black market survey, I obtained virtual copies of 35 darknet black market forums spanning over a period of 4 years [8].

2.1 Market crawling

A large part of the work was to create a crawler to extract relevant data from these full darknet pages. This was done using python - the crawler was programmed to analyse each page of each market and to locate user discussions within the forums and log each post of each user with the corresponding topic and each occurrence of inter-user communication both timestamped. This allowed me to create two initial edge lists for each of the black markets - one linking each user with each topic they participate in, and the other linking the users with each other in the cases when they communicate directly with each other.

2.2 Forum overview

Out of the 35 darknet black markets I further analysed 26 of them. 9 markets were omitted due to either a lack of quality in the data, like for example when it was obvious that parts of the original forums are missing in the scraped pages, or because they were too small for a meaningful network structure analysis. The data obtained is over a 4 year period spanning from July 2011 to July 2015. Figure 2.1 illustrates each of the markets activity period which was analysed. This almost perfectly corresponds to the real time period in which the market was active. From creation, to either being shut down by law enforcement or by the market itself performing a so called ”Exit scam”,

in which the operators of the marketplace shut the market down by themselves, stealing all of the cryptocurrency held on the sites virtual wallets by vendors and buyers (this was the case for the black market Evolution for example). The only exceptions are those of the 5 markets who terminate on July 2015, simply because that was the last date of the information sources [8] performed scrape. Out of these five, two markets called ”TheHub” and ”Agora”, remain active at the time of writing this thesis alongside approximately 20 new ones that have formed since then .

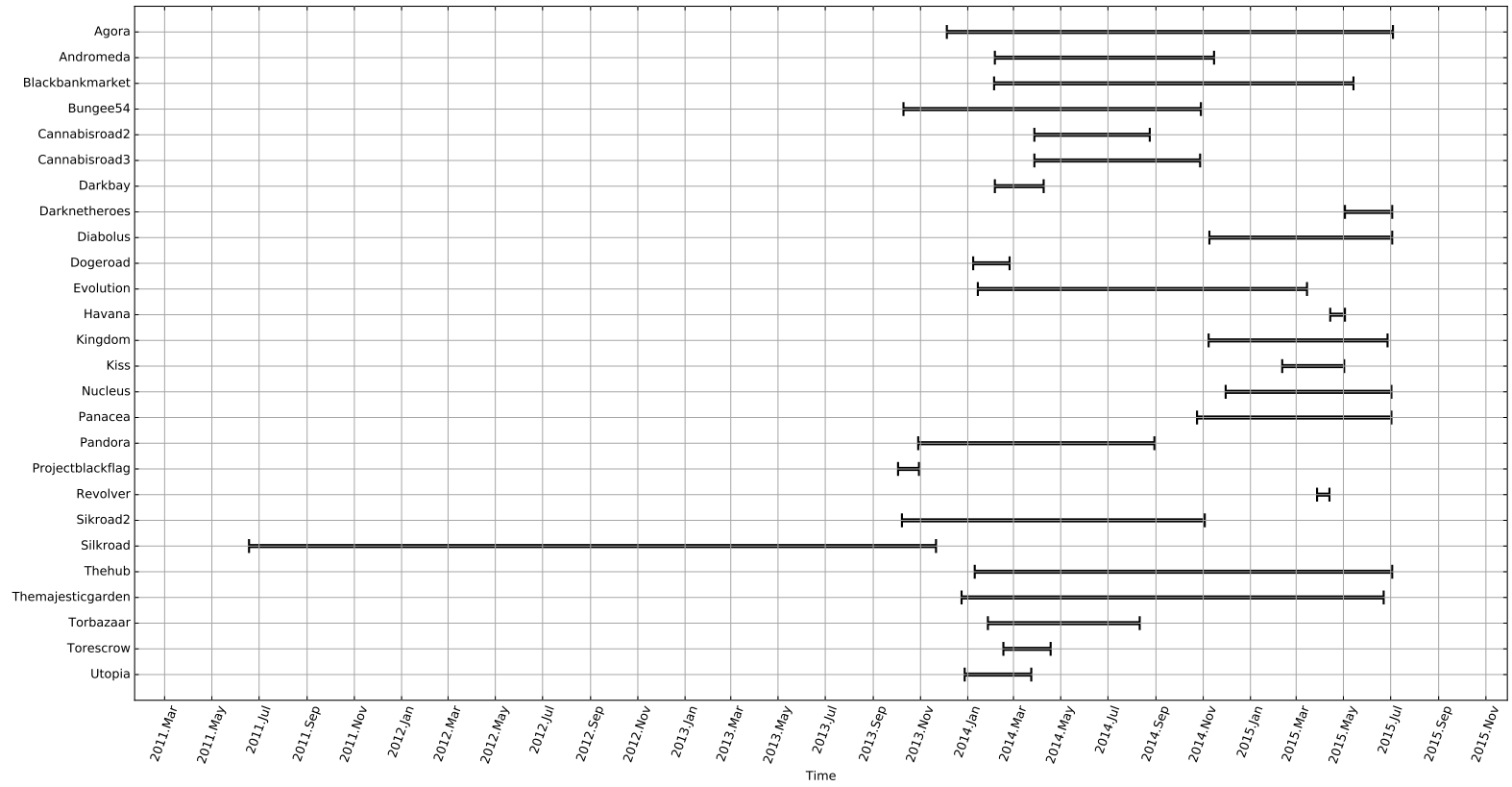

Figure 2.1: *Time span of darknet black market activity period*

On a side note - this graph also clearly illustrates the evolution of popularity of darknet black markets. When the first marketplace, the original ”Silkroad” was active for 2 years, it was probably the only one out there, and it’s popularity was not as wide spread as the concept of darknet black markets is today. After its shut-down by the FBI, and the giant commotion led by the media following the court case of its alleged owner Ross Ulbricht, this type of illicit activity on the darknet became increasingly popular.

2.3 Data extraction

To extract and analyse the data of the network structure of these black market forums, I built a crawler program in python. The crawler analysed the forums page by page, searching for users and looking at communication between them. When parsing the data in a graph structure, the users are represented as nodes, and an edge is drawn between them, whenever a communication occurs. Also other types of representation of the network structures were analysed as described in the following subsections.

Figure 2.2 illustrates the relative sizes of the markets by the number of their active users. This is the number of nodes in each network. Figure 2.3 illustrates the occurrences of communication between users in each market. This is the number of edges in each network.

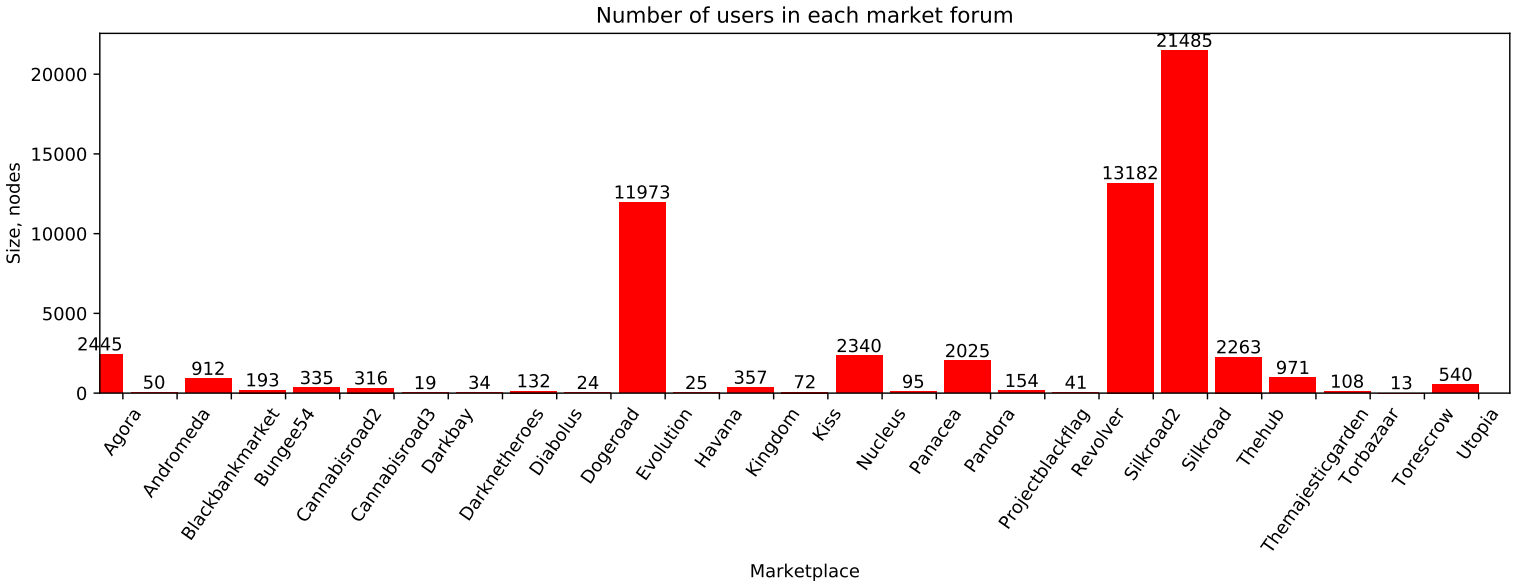

Figure 2.2: *Active users or nodes in each of the markets*

The following two subsections explain in depth how the network representations of these darknet communities were created.

2.3.1 Communication networks

The communication network is a directed graph of the explicit communications between forum users. In online forums users have the possibility to quote one another. Therefore each time a citation is found, a directed edge pointing from the user who cites to the user who is cited is added. Usually when someone cites someone else in

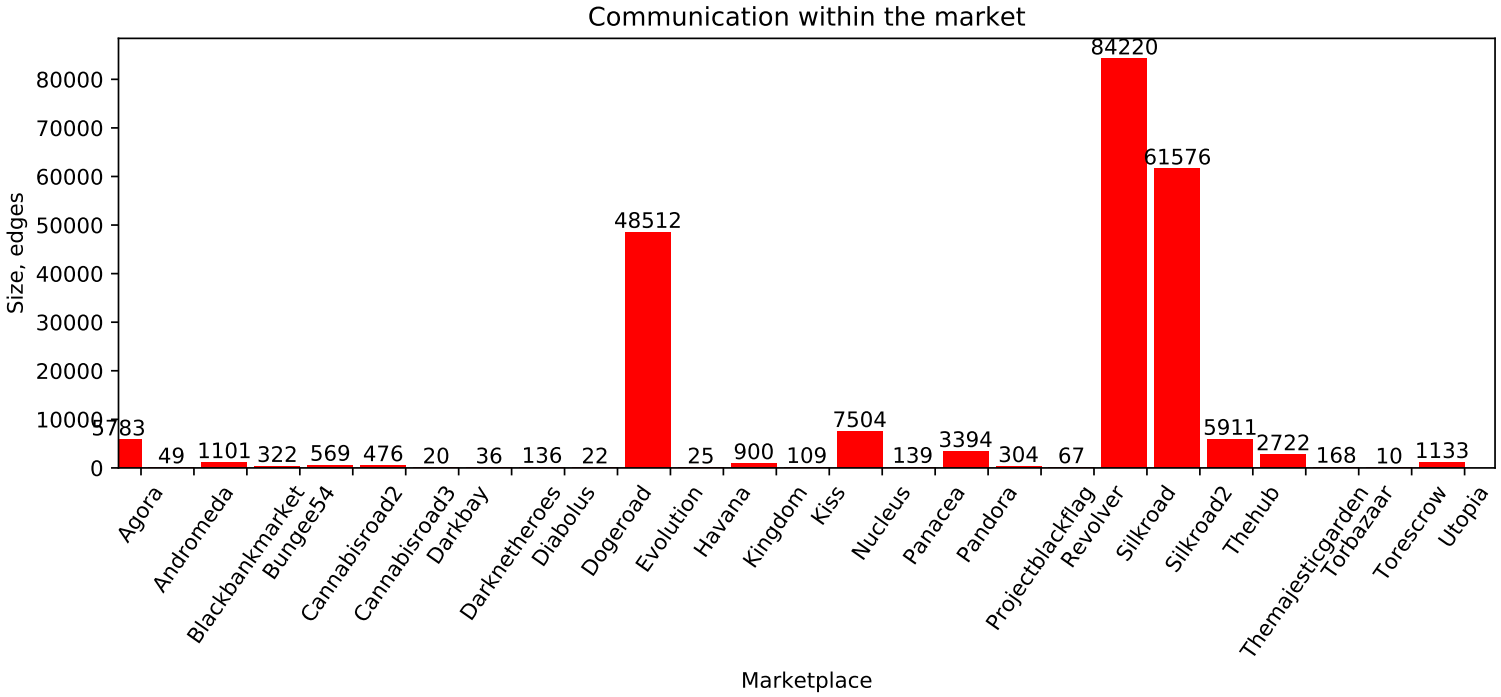

Figure 2.3: *Communication occurrences between users - the number of edges in each market*

these forums, it is followed by a response to the user that wrote the original message. Therefore these citing cases can be regarded as events of communication and are added in the graph. In Figure 2.4 a smaller communication network called cannabisroad2 is shown for illustrative purposes.

The disadvantage of this data extraction method is that the edges might be underestimated. The reason for this is the fact that not all users use the 'quote' option to talk directly to each other. Some people tend to simply address the username (in many cases abbreviating it in a random manner, which prevents automatic extraction of these communications) of the person to whom they wish to speak to, and then carry on with their message. Never the less the fraction of users that do use the 'cite' or 'quote' option in these forums is significant enough to draw conclusions of how this communication network is structured. However, to some-what mitigate the influence of this effect, and broaden the analysis of these networks, another type of data extraction method was used to generate expertise networks of the same markets.

2.3.2 Expertise networks

The Expertise network method equivalent to that used in a publication by Jun Zhang et al. in "Expertise Networks in Online Communities: Structure and Algorithms" [18].

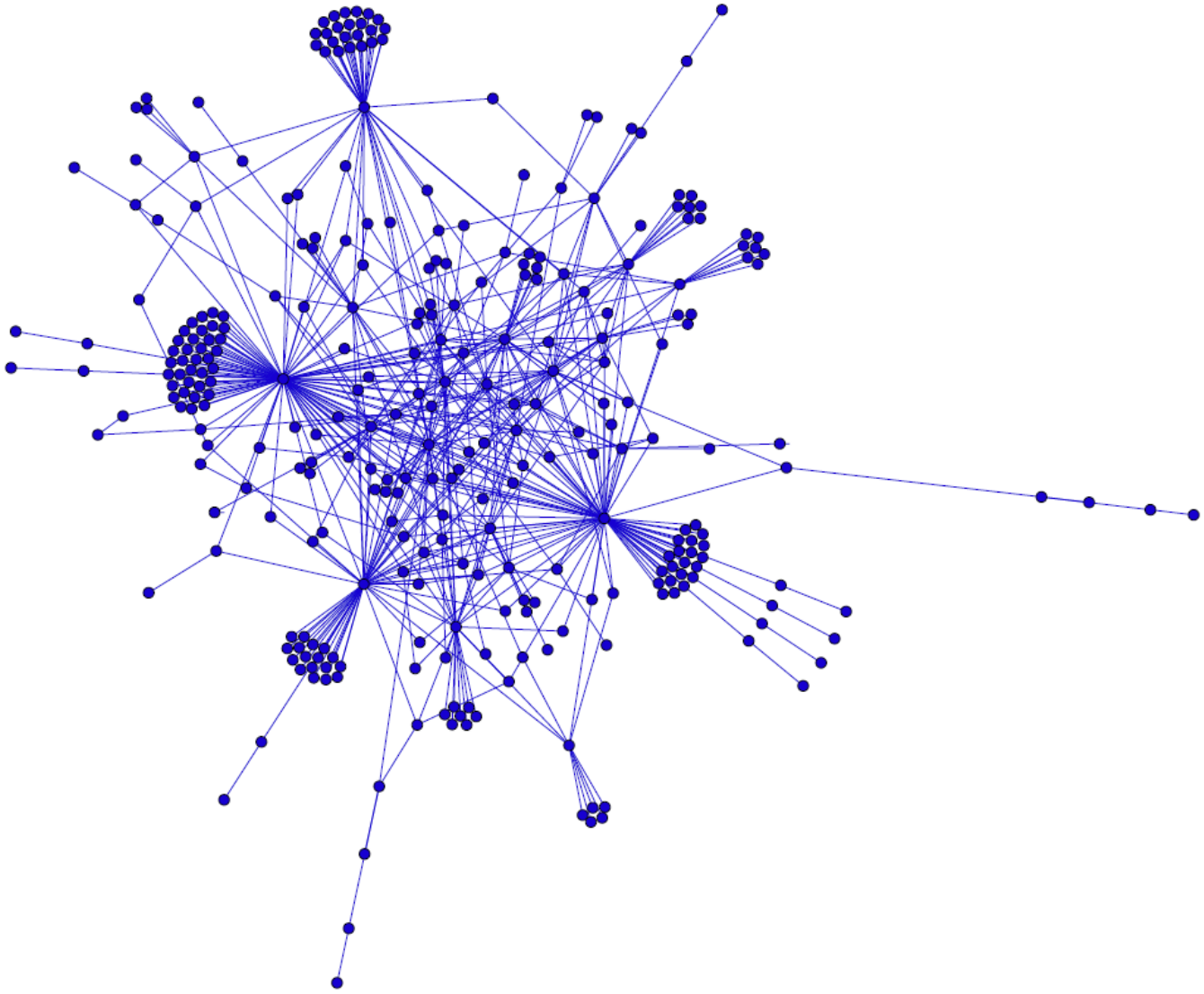

Figure 2.4: *cannabisroad2 communication network*

This is based on the idea that people who usually create topics tend to either ask a question abut a topic to all of the forum members, or give an opinion on a subject provoking a follow-up discussion between the members of the forum. This then means, that all of the people participating in a topic share expertise on the subject that is being discussed. Therefore in this method directed links are drawn in graph from each user that replies to a topic to the user who originally posted the topic.

Naturally the case investigated in this thesis differs from the work done by Zhang et al. A darknet black market forum does not only consist threads aimed to ask questions about a certain topic, as they do in the online communities (mostly centred around the topic of programming) explored in [18]. Therefore it is appropriate to conclude that this approach will lead to overestimation of the edges of the network. The main focus

of this work is to analyse the Communication networks in depth. Therefore the Expertise networks will only serve as additional data to support or dispute any conclusions made about these 25 darknet black market forum Communication networks.

However for a deeper understanding of the topological properties of these networks a different projection of the expertise networks is also analysed. In this projection the network is constructed as a bipartite graph with one layer being the users and the other - the topics they take part in. The user nodes are connected to those topic nodes that they have taken part in. This allows for a deeper analysis of the expertise network, and serves as an additional argument for the conclusions later drawn about the communication networks.

Chapter 3

Basic complex network properties

This chapter investigates the static complex network properties of darknet black market graphs.

3.1 Network types

Complex networks are usually classified as being either binary - meaning that either there is or is not an edge between any two nodes, or weighted - meaning that the edge carries extra information about it. This extra information is usually called its strength, which in this case might be the amount of communication occurrences between the nodes. However the edge strength is out of the scope of this research.

Graphs are also usually classified as either directed or undirected. Both cases are analysed in this work - in the directed case the edges of the network are drawn from the user which quotes another user pointing to that user who is quoted.

3.2 Degree

One of the simplest ways to characterise a node in a network is its degree, which measures the number of connections between the node and other nodes. For directed networks, like the communication and expertise networks analysed in my work, a nodes degree can be broken down into two distinct categories: the out-degree - the number of directional ties emanating from it, and the in-degree - the number of directional links that point towards it. Usually in directed networks that model communications between people, the nodes in-degree signifies its popularity while the out degree can be regarded as expansiveness[11]. In my analysis, for the Expertise networks, a forum

users in-degree is the sum of all individual forum users that have replied to a topic started by that user, and the out-degree is the number of unique user's topics to which that user has replied. In the communication network the out-degree of a user shows how many unique user's he or she has quoted, and the in-degree shows how many unique users have quoted that user.

3.3 Power-law degree distributions

Once the degree of all nodes of the graph is calculated, it is natural to ask how big of a fraction of the vertices have a certain degree. This is known as the degree distribution of the graph. The degree distribution as the degree itself can also be broken down into the in-degree and the out-degree distributions. Most large real world networks display an interesting feature - the vertex connectivities follow a scale-free power-law distribution [3]. A power law is a function that decreases as it's argument to some fixed power [7]. In the case of networks, this is usually written as

$$p(k) \sim k^{-\gamma}$$

Where γ is the fixed exponent characterising how fast does the number of nodes decay with increasing degree. A larger γ would signify a steeper slope and therefore less nodes with higher degree value.

Network type	γ
WWW[in]	2.00
WWW[out]	2.31
Mobile phone calls[in]	4.69
Mobile phone calls[out]	5.01
E-mail[in]	3.43
E-mail[out]	2.03
Science collaboration	3.35
Actor collaboration	2.12
Citation network[in]	2.79
Citation network[out]	4.00

Table 3.1: *γ values for some real world networks [2]*

Network type	γ
Orkut social network	0.747
LiveJournal blogging community	1.032
Wikipedia author network	1.95
YouTube social network	1.42
DBLP author network	1.205
SlashDot user community network	1.215
Enron e-mail communication network	1.164

Table 3.2: *γ values for some real world online networks [9]*

Studies of different types of networks reveal different values for γ. Table 3.1 lists a few of them. This list shows that in these cases the value of the exponent $\gamma \geq 2$, however following the research done in [9], where the investigated networks are online-based communities values of $\gamma < 2$ seem to appear. These results are summarized in Table 3.2.

In my own analysis for the darknet black market forum networks, the results for the exponent γ reflect more those in Table 3.1, than in Table 3.2. Typically the exponents of the communication network in-degree distributions lie between 1.35 and 1.75, with a single exception of a black market called "Blackbankmarket". The potential reasons for this exception will be analysed afterwards. The out-degree exponents of communication networks lie between 1.3 and 1.65 with no exception. A typical mid-sized market's called "The Hub" forum communication network in and out-degree distributions are shown in Figure 3.1 with their corresponding power law fits.

Similarly if we look at the degree distributions for the expertise networks we find that for in-degree distributions $\gamma \in [0.8; 1.17]$ and for out-degree distributions $\gamma \in [1.39; 1.92]$. An example of an expertise networks degree distribution is shown in Figure 3.2

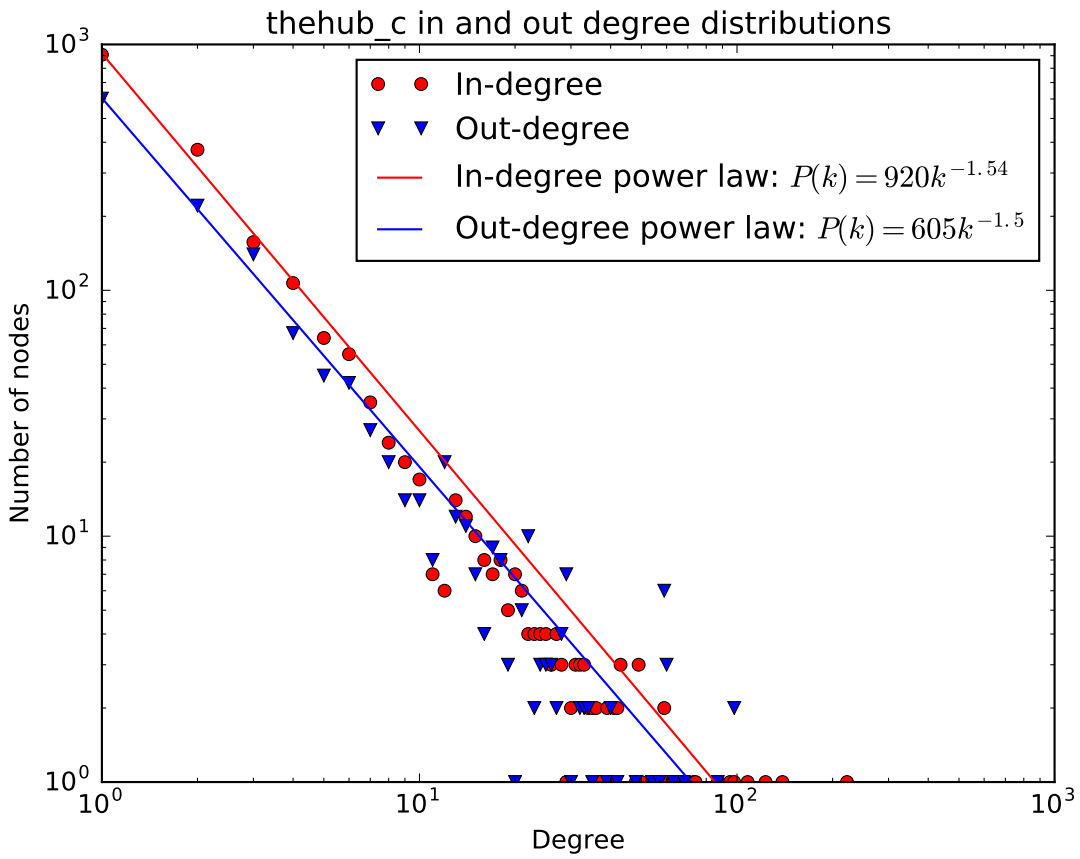

Figure 3.1: *Degree distribution of The Hub communication network*

3.4 Assortativity

Another interesting network property to look at is it's degree correlations. These are calculated with the average nearest neighbor degree defined as

$$k_i^{nn} = \frac{\sum_{j \neq i} \sum_{k \neq j} a_{ij}^* a_{jk}^*}{\sum_{j \neq i} a_{ij}^*}$$

where a_{ij} is the entry of the i'th row and the j'th column of the networks adjacancy matrix. One could also ask the question whether on avaerage there is a correlation between the nodes degree and the degree of it's neighbours. A more rigorous definition of this property is given by M. E. J. Newman and Juyong Park in [14]. If p_k is the degree distribution of our network (the fraction of vertices with degree k in the network), they define the properly normalized distribution of the excess degree as

$$q_k = \frac{(k+1)p_{k+1}}{\sum_k k p_k}$$

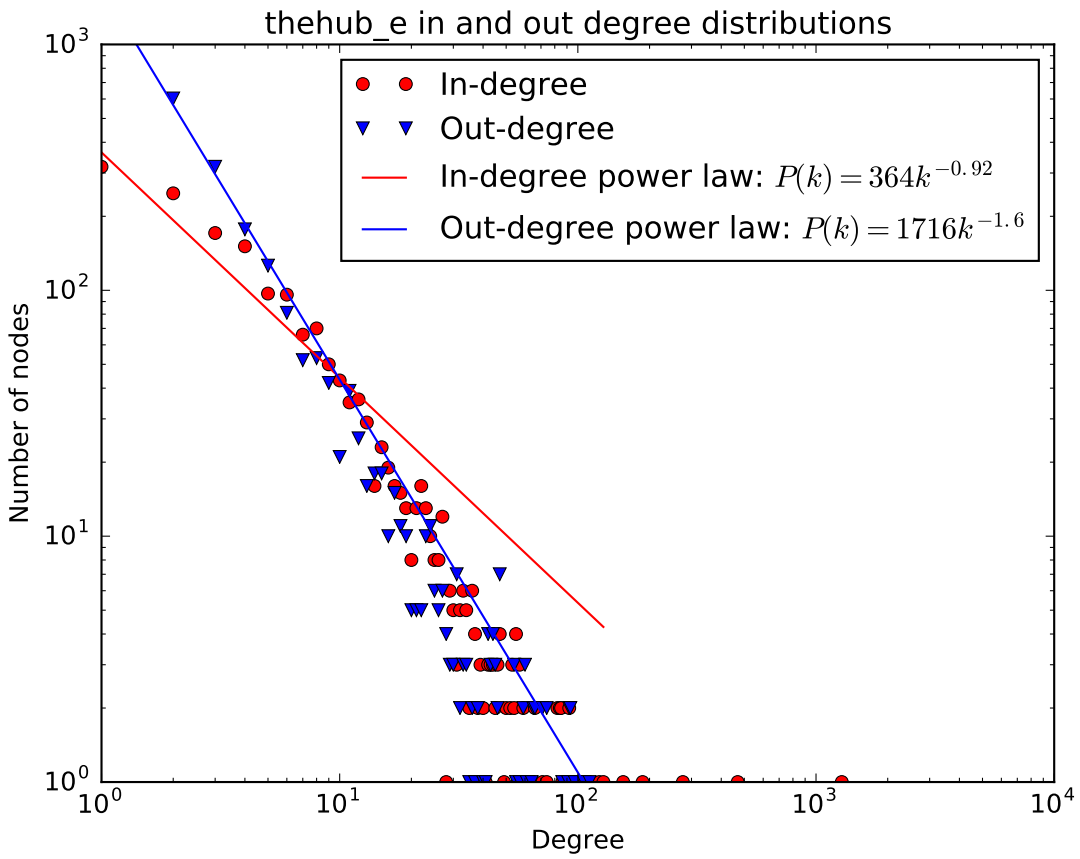

Figure 3.2: *Degree distribution of The Hub expertise network*

and the degree assortativity coefficient as

$$r = \frac{1}{\sigma_q^2} \sum_{jk} jk(e_{jk} - q_j q_k)$$

where e_{jk} is the joint probability that an edge chosen at random will have vertices with excess degrees j and k at its endpoints, and σ_q^2 is the variance of q_k. When $r > 0$ the network is said to be assortative and when $r < 0$ it is disassortative.

The works of Newman [12] and [13] show that social networks like science co-authorship, film actor collaboration, and email address book networks all exhibit assortative mixing. This means that nodes that have many connections tend to be connected to other nodes with many connections. More recent works like [5] have shown that the same is true for the modern and hugely popular online social networks like Twitter, Facebook, Google+, LinkedIn and others. However, upon the analysis of the darknet forum structure, a completely different situation emerges - almost all show signs of dissassortative mixing, meaning that nodes with a high degree have a bias of connecting to nodes with a smaller degree. This is shown in table 3.3

Market	Network type	Assortativity coefficient
abraxas	Communication	-0.128777238152
abraxas	Expertise	-0.2511857585
blackbankmarket	Communication	-0.133875488198
blackbankmarket	Expertise	-0.185861442577
evolution	Communication	-0.0606666840915
evolution	Expertise	-0.139172617107
nucleus	Communication	-0.0457730106701
nucleus	Expertise	-0.178889540218
pandora	Communication	-0.0603477830086
pandora	Expertise	-0.107607117244
silkroad1	Communication	-0.0485285823462
silkroad1	Expertise	-0.0901921852604
thehub	Communication	-0.145032160299
thehub	Expertise	-0.169944328433
themajesticgarden	Communication	-0.164333213271
themajesticgarden	Expertise	-0.275730373008
torbazaar	Expertise	-0.336354922087
utopia	Communication	-0.215538949141
utopia	Expertise	-0.228836887026

Table 3.3: *Darknet black market assortativity*

3.5 Clustering

The clustering coefficient is a third order network property. There exists a single definition for the undirected clustering coefficient, but in the directed case there are several ways to define it.

3.5.1 Undirected case

For the undirected graphs the clustering coefficient can be defined globally like in [14] as the probability, averaged over the network, that two of a nodes neighbours will have an edge drawn between them. This is obtained by first counting the number of triangles (sets of 3 vertices that are all connected to each other) and wedges (sets of vertices that are connected to an unordered pair of others). Then the clustering coefficient or graph density is defined as:

$$C = \frac{3 \times \text{number of triangles on the graph}}{\text{number of wedges}}$$

For a single node the clustering coefficient c_i is the number of connections between neighbours of node i divided by the the number of connections possible:

$$c_i = \frac{\sum_{j \neq i} \sum_{k \neq i,j} a^*_{ij} a^*_{jk} a^*_{ki}}{\sum_{j \neq i} \sum_{k \neq i,j} a^*_{ij} a^*_{ki}}$$

A high clustering coefficient means that more tightly knit communities are present in the network characterised by a high density of links such that their likelihood is larger than the average probability of a link randomly established between two nodes [17].

3.5.2 Directed case

For the directed case in my work I define two different clustering coefficients for node i as

$$c_i^{out} = \frac{\sum_{j \neq i} \sum_{k \neq i,j} a_{ij} a_{ik} (a_{kj} + a_{jk})}{\sum_{j \neq i} \sum_{k \neq i,j} a_{ij} a_{ik}}$$

$$c_i^{in} = \frac{\sum_{j \neq i} \sum_{k \neq i,j} a_{ji} a_{ki} (a_{kj} + a_{jk})}{\sum_{j \neq i} \sum_{k \neq i,j} a_{ji} a_{ki}}$$

Version of September 18, 2017– Created September 18, 2017 - 14:03

where c_i^{out} is the number of connections between out-neighbours of node i (the connections which point from i to its neighbours) divided by the number of out-neighbours possible. Similarly c_i^{in} is has the same definition with in-neighbours.

3.6 Time stamps

Because the data obtained from the darknet networks is timestamped - it is possible to know the exact time each event of communication occurred, it is possible to observe how the network growth happened over time. This is used for analysis in later chapters to see whether certain network properties are present from the beginning or do they develop over time as the network grows. This allows for a more dynamic view of the network, rather than just the static final image.

Chapter 4

Pattern detection methods

This chapter describes the Maximum Likelihood method [15] used to investigate the topological properties in order to determine patterns in the networks that were analysed in this work.

4.1 Analytical maximum-likelihood method

The paper [15] published by Tiziano Squartini and Diego Garlaschelli introduces a novel method for analytically obtaining expectation values of any topological property for any binary, weighted, directed or undirected network. This thesis uses their methods for both undirected and directed binary networks to asses whether the networks higher order properties, like the average nearest neighbour degree or the clustering coefficient, arises because of low level constraints, or are there other structural patterns present. Specifically comparing the darknet black market forum graphs with the Configuration model, which contributes an ensemble of random networks having, on average, the same degree sequence as the real network. In essence this amounts to asking whether the sole variable describing these networks is the amount of communication between users, and everything else is as good as random, or are there other patterns in the user behaviour.

4.1.1 Undirected networks

Their paper shows that in order to do so for the undirected case, one must first find a N-dimensional vector $\vec{x} = \{x_1, ..., x_N\}$ of parameters, which can be found by solving a

set of N coupled non-linear equations

$$\sum_{j \neq i} \frac{x_i^* x_j^*}{1 + x_i^* x_j^*} = k_i(\mathbf{A^*}) \quad \forall i$$

where $k_i(\mathbf{A^*})$ is the observed degree of vertex i in the real network $\mathbf{A^*}$. These parameters in turn allow us to determine the expectation values of the adjacency matrix entries (which are also equal to the probability of there being an edge between nodes i and j).

$$p_{ij}^* = \langle a_{ij} \rangle^* = \frac{x_i^* x_j^*}{1 + x_i^* x_j^*}$$

4.1.2 Directed networks

As for the directed case one must find two N-dimensional vectors $\vec{x}$ and $\vec{y}$ of parameters that solve the following set of 2N coupled non-linear equations:

$$\sum_{j \neq i} \frac{x_i^* y_j^*}{1 + x_i^* y_j^*} = k_i^{out}(\mathbf{A^*}) \quad \forall i$$

,

$$\sum_{j \neq i} \frac{x_j^* y_i^*}{1 + x_j^* y_i^*} = k_i^{out}(\mathbf{A^*}) \quad \forall i$$

where $k_i^{out}(\mathbf{A^*})$ is the observed out-degree and $k_i^{in}(\mathbf{A^*})$ is the observed in-degree of vertex i in the real network $\mathbf{A^*}$. This again allows to obtain the expectation values of the adjacency matrix.

$$p_{ij}^* = \langle a_{ij} \rangle^* = \frac{x_i^* y_j^*}{1 + x_i^* y_j^*}$$

4.1.3 Bipartite networks

The directed case can also be applied to study undirected bipartite networks in the same manner, so that additional conclusions can be made from the Expertise networks. To do this one must simply project the bipartite network onto a directed binary network with each of the parts having only the in-degree or out-degree respectively. Then the same expression for the expectation values as in the directed case can be used.

Once these expectation values are obtained, they can be used to calculate any topological property that is defined by an expression that incorporates the values of the networks' adjacency matrix. To do so we must simply replace these values a_{ij} with their expectation values $\langle a_{ij} \rangle$. In this thesis this approach is used to calculate for each of the black market networks' each nodes' expected value for both nearest neighbour

degree and clustering coefficient. Then these expected values are averaged over all the nodes with the same degree and they are compared to the average real values of the network to see whether there are any discrepancies.

4.2 Unbiased sampling method

However, not all of the topological properties of interest to my work can be defined by expressions consisting of only the values of the networks' adjacency matrix. And even if they could be in principle, in order to check some minor hypothesis along the way, a less time consuming method is also used. This method is published by Tiziano Squartini, Rossana Mastrandrea and Diego Garlaschelli in [16]. It uses the results of [15] to generate unbiased samples of networks where the constraints (the degree sequence in this case) are realized as ensemble averages. To do so, the same expected values for the adjacency matrix $\langle a_{ij} \rangle$ described in the previous paragraphs are utilized, but in this case I make use of the fact that they are also equal to the probability p_{ij} of there being an edge between nodes i and j. This in turn allows me to quickly and efficiently generate many network samples drawn from maximum-entropy distributions, and investigate these by averaging over them any property of interest. An additional benefit of using two methods is that this also allows me to check for errors in seeing whether the results of both of the methods match up to each other.

Version of September 18, 2017– Created September 18, 2017 - 14:03

Chapter 5

Results

In this section I describe the results obtained by using the methods from the previous chapter to analyse of the topology of darknet black market networks.

5.1 Analytical approach

To begin with all of the darknet market networks are investigated with the method described in section 4.1. The resulting average values of the topological properties are plotted as a function of the node degree.

5.1.1 Nearest neighbour degree

The first investigated topological property is the average nearest neighbour degree defined in Section 3.4. Figures 5.1, 5.2, 5.3 and 5.4 show it as a function of the node degree for each of the darknet black market networks. It is visible that for most networks the theoretical values roughly match with the values of the real network. The dispersion of the real values might be larger then expected, this is further analysed in the following sections. It is worth mentioning that the spread of the real values around the theoretical curve seems not to be symmetric - the real values more often then not fall below the theoretical curve. This points to an inference that these networks exhibit on average a slightly lower nearest neighbour degree than the Configuration Model predicts they should. The significance of this deviation is further analysed in the coming sections. The only two networks that cardinally disagree with the theoretical values are also amongst the smallest - andromeda and darknetheroes.

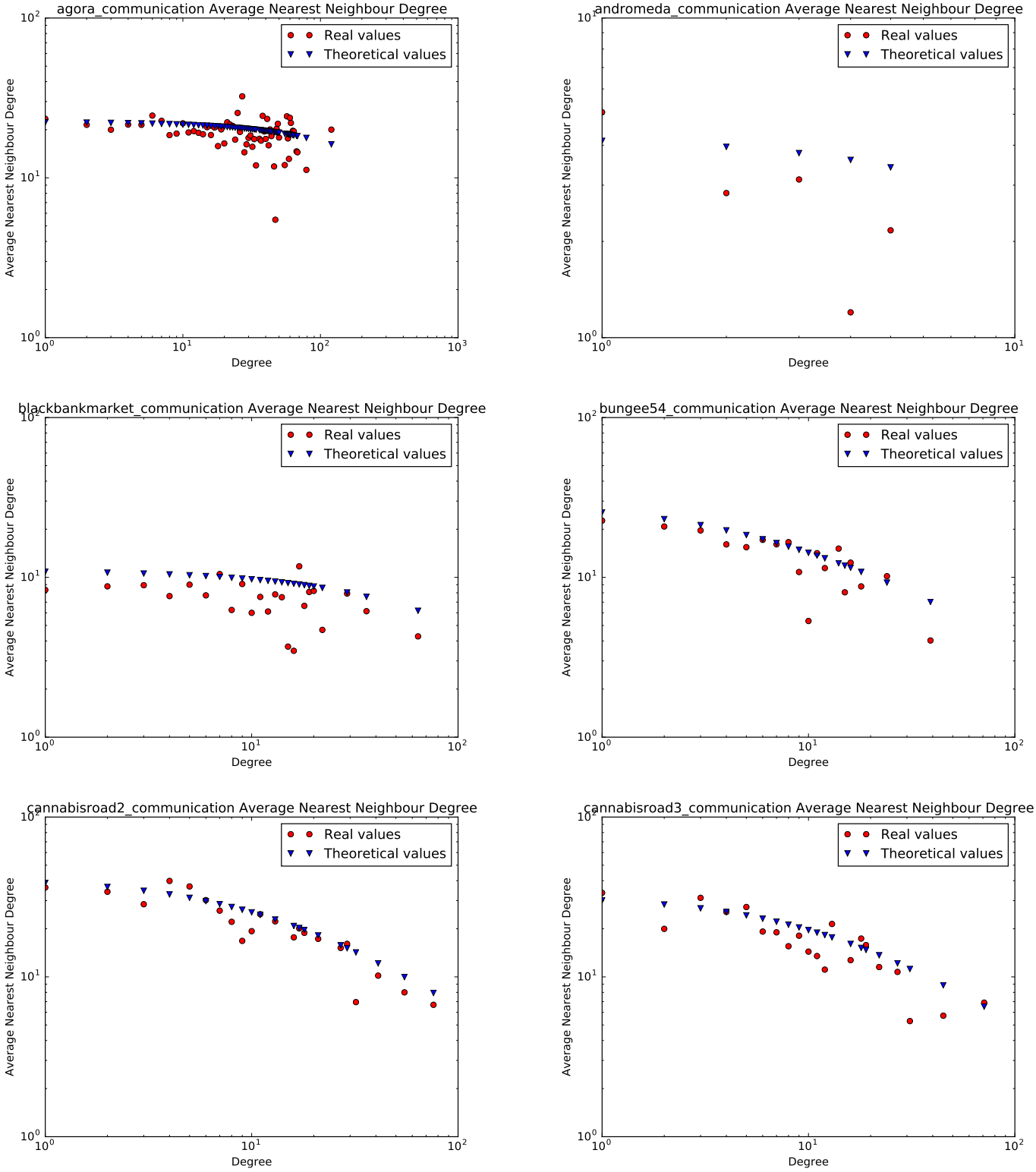

Figure 5.1: *Plots of average nearest neighbor degrees*

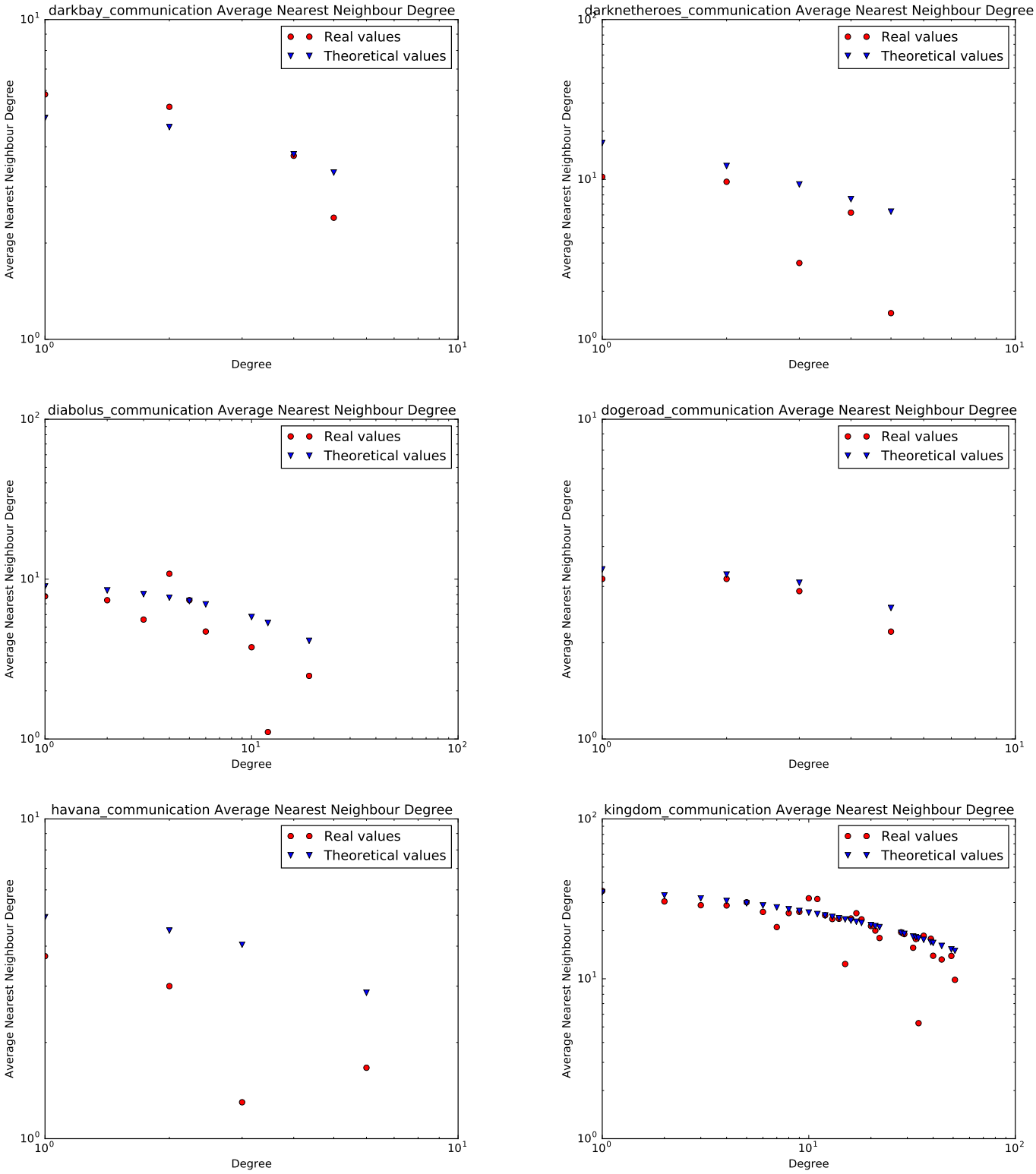

Figure 5.2: *Plots of average nearest neighbor degrees*

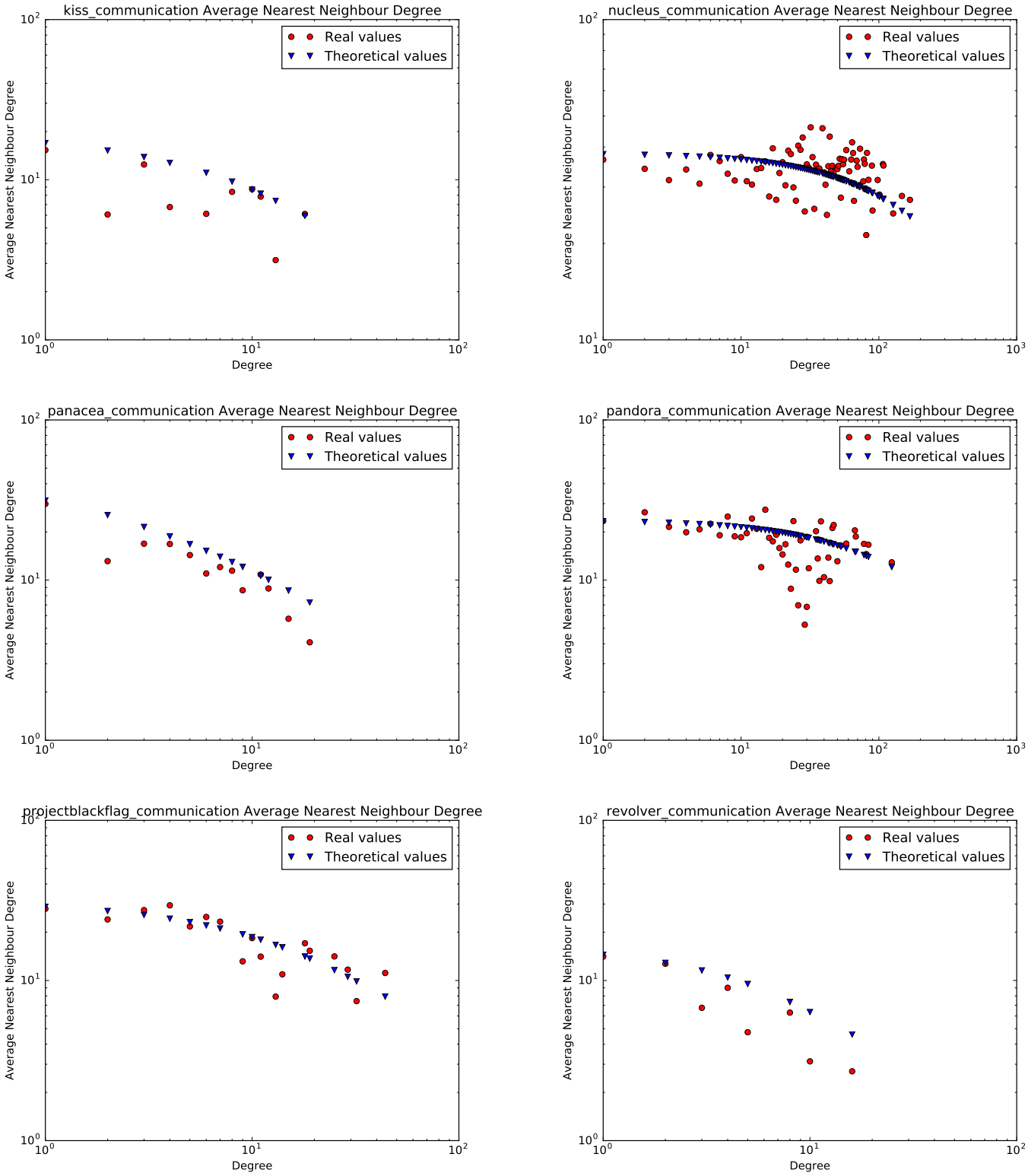

Figure 5.3: *Plots of average nearest neighbor degrees*

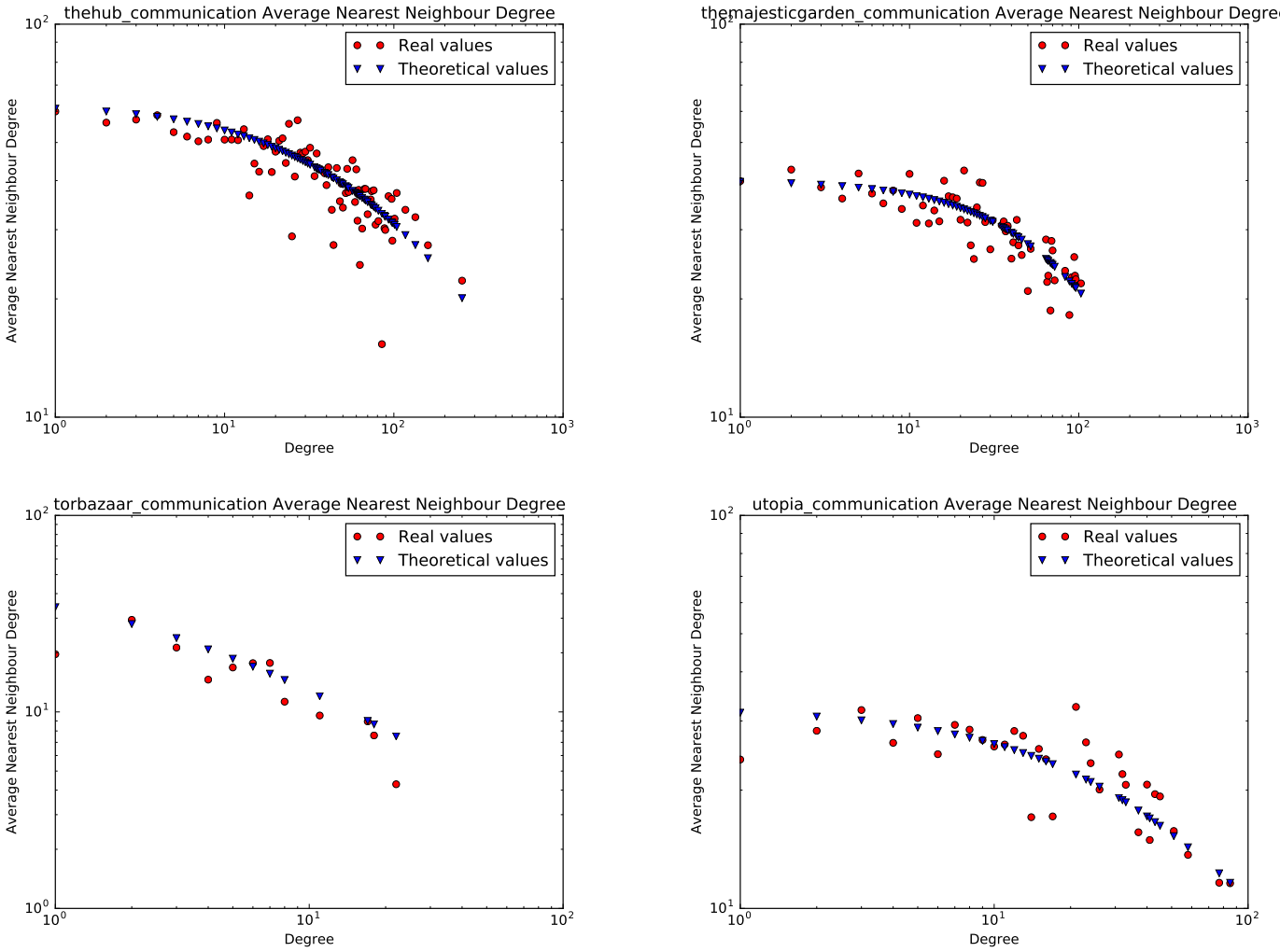

Figure 5.4: *Plots of average nearest neighbor degrees*

5.1.2 Clustering coefficient

The second topological property investigated with the maximum likelihood method is the clustering coefficient defined in Section 3.5. Figures 5.5, 5.6, 5.7 and 5.8 show it as a function of the node degree for each of the darknet black market networks. These graphs reveal an entirely different scene than the preceding ones that looked at the average nearest neighbour degree.

Firstly the reader might notice that some of the networks, namely: darkbay, darknetheroes, dogeroad, torescrow - have been omitted here. This is due to the fact that those networks are too small to produce the clustering coefficient curves worth analysing. As a rule of thumb in this research, if a network can not produce at least 3 points on the real value graph, it would not be justified to draw any conclusions from it.

Secondly only few of the results can be regarded as somewhat matching the values predicted by the Configuration Model. Amongst these I would consider cannabisroad2, cannabisroad3, kingdom, projectblackflag, torbazaar and utopia. How good exactly is this match will be investigated in further Sections. This already shows that the Configuration Model is not sufficient in explaining the higher order properties of these networks. This point is further examined in Section 6.2

And finally amongst the networks that clearly do not reflect their expected clustering coefficients in the majority of cases this mismatch reveals itself in a systematic manner. This systematic mismatch is best observed in the larger networks like agora, nucleus, pandora, and thehub. In all of these cases the real values form a correlation that is steeper than the theoretical curve. It starts off in the small degree range with values exceeding those that the Configuration Model predicts, then intersects the theoretical curve and subsequently some of the largest degree nodes either fall on the theoretical curve or below it. This means that nodes with a small degree are part of way more triangles than the theory predicts, and only those with a relatively high degree exhibit a behaviour that can be described by Configuration Model. Possible interpretations of this phenomena are examined in the discussion and conclusions of this work

Version of September 18, 2017– Created September 18, 2017 - 14:03

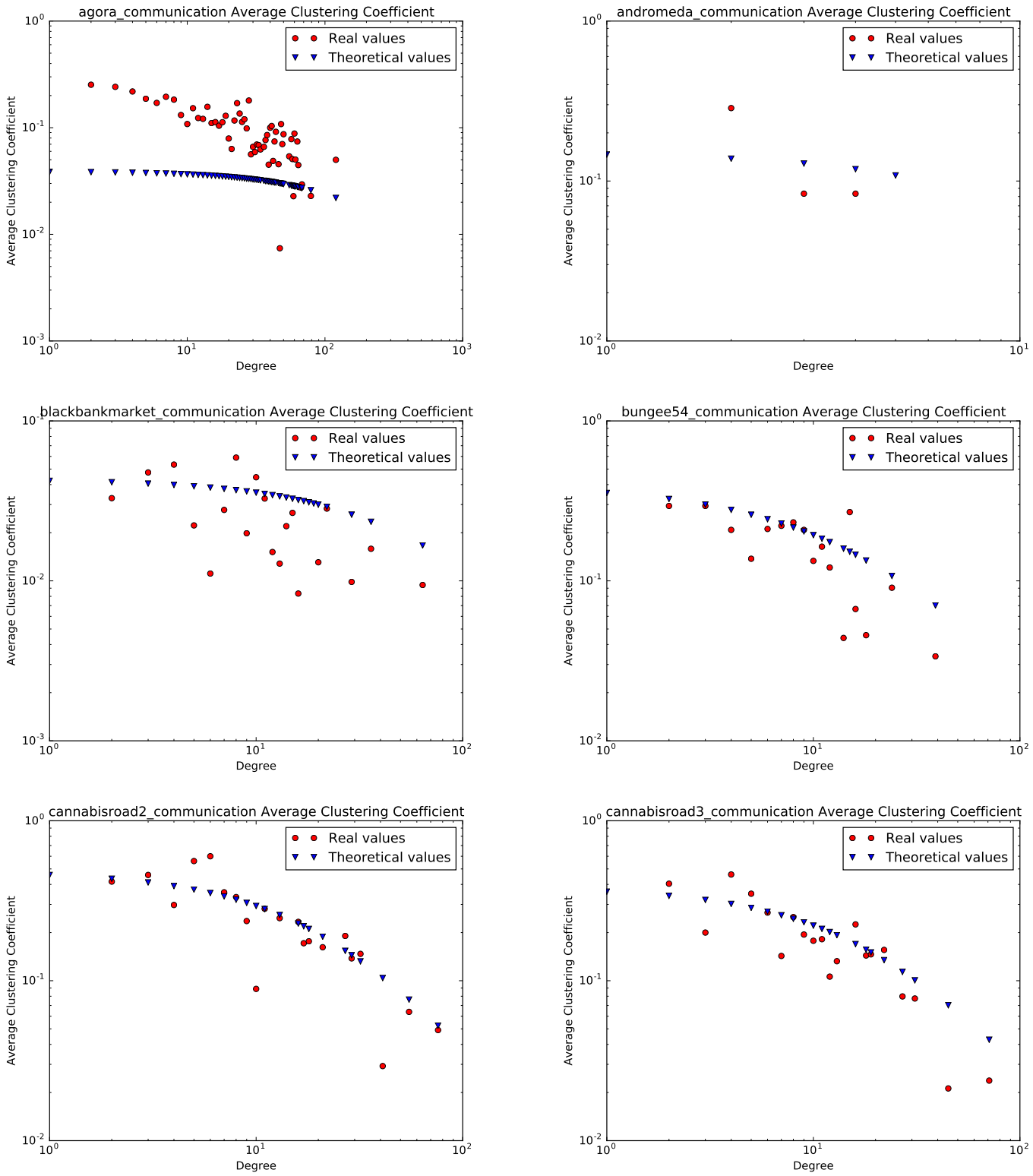

Figure 5.5: *Plots of average clustering coefficients*

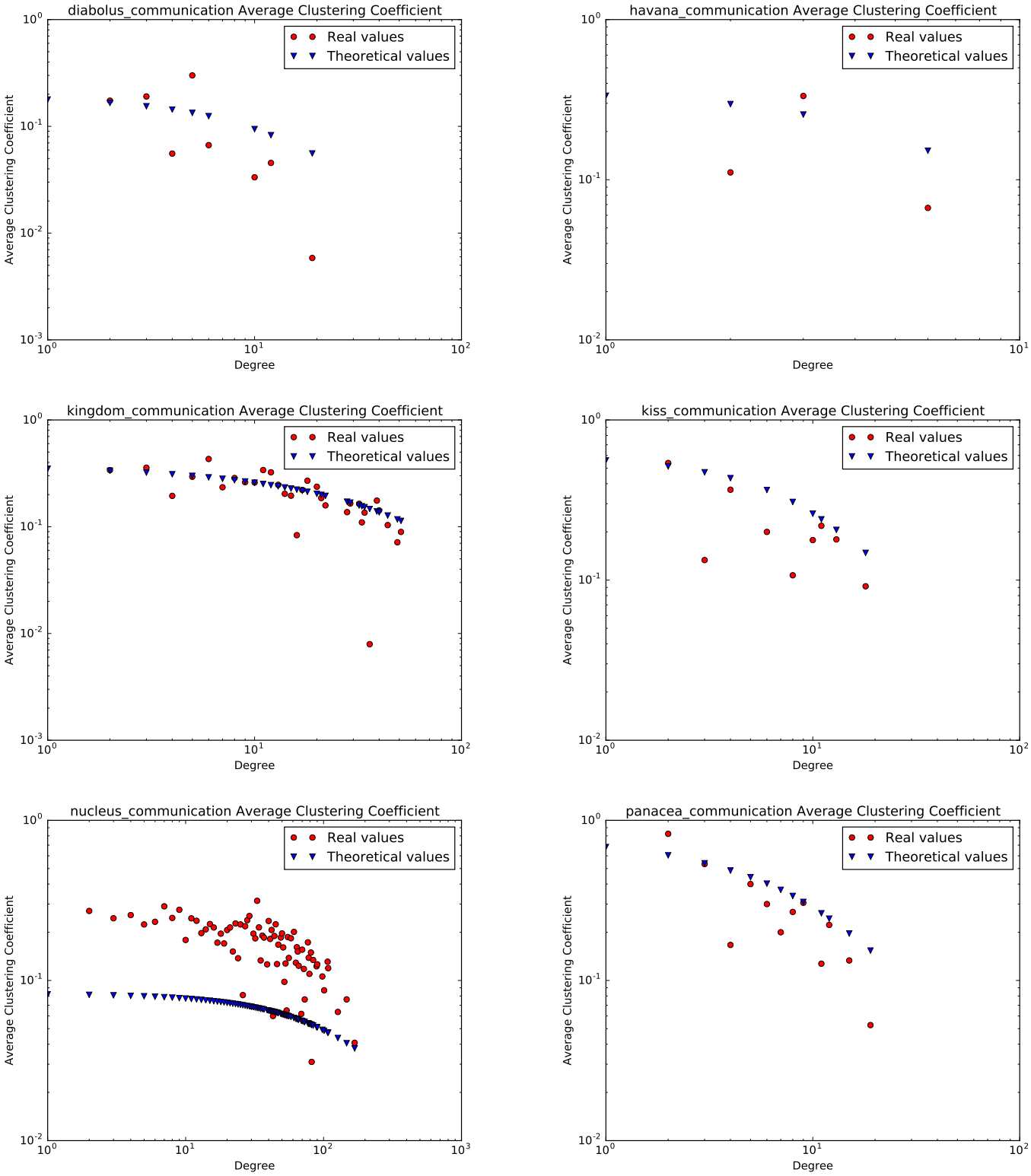

Figure 5.6: *Plots of average clustering coefficients*

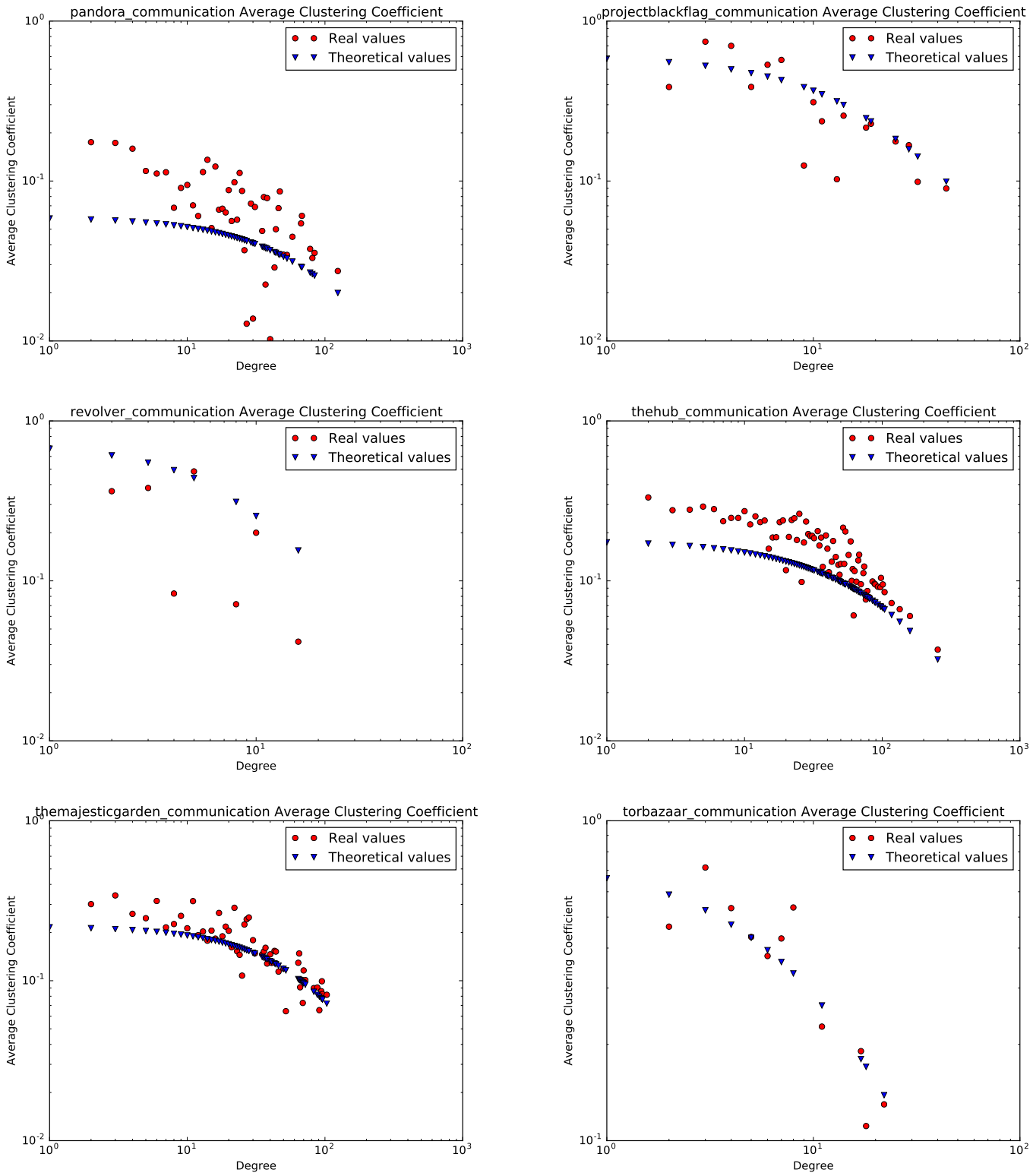

Figure 5.7: *Plots of average clustering coefficients*

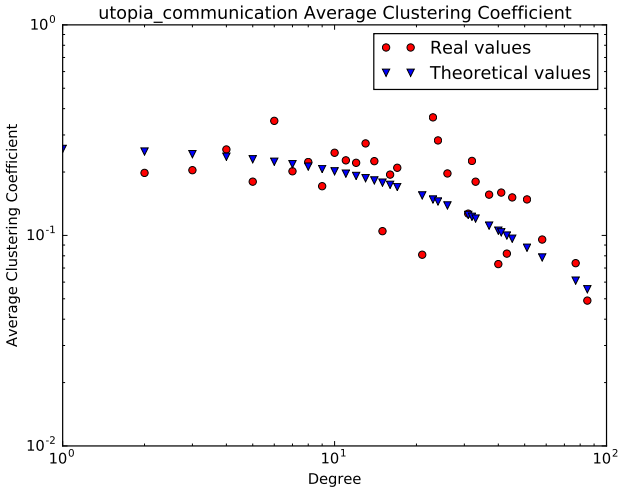

Figure 5.8: *Plot of average clustering coefficients*

5.2 Sampling approach

This Section further explores the analysis of the darknet black market networks started in the preceding section. It uses the method described in Section 4.2 to first check the results of the previous method and second - analyse and quantify the features of the graphs that are not so easily defined in analytical expressions involving their adjacency matrices.

In the coming two subsections the results are shown for a set of 6 different darknet black market networks. These are chosen from different size scopes and different degrees of accordance with the Configuration Model predictions based on the results described in the preceding two sections. The sampled 6 network ensembles are agora, andromeda, cannabisroad3, pandora, thehub, themajesticgarden.

5.2.1 Nearest neighbour degree

First we look at how well actually does the Configuration Model predict the average nearest neighbour degree for these networks. Each of the networks have been sampled 200 times using the method from Section 4.2. For each degree value the mean and standard deviation of the average nearest neighbour degree is calculated over the whole ensemble. These results are shown in Figure 5.9. The *Pr* values on each graph show the fraction of values from the real network found within 1, 2 and 3 standard deviations from the ensemble averages respectively.

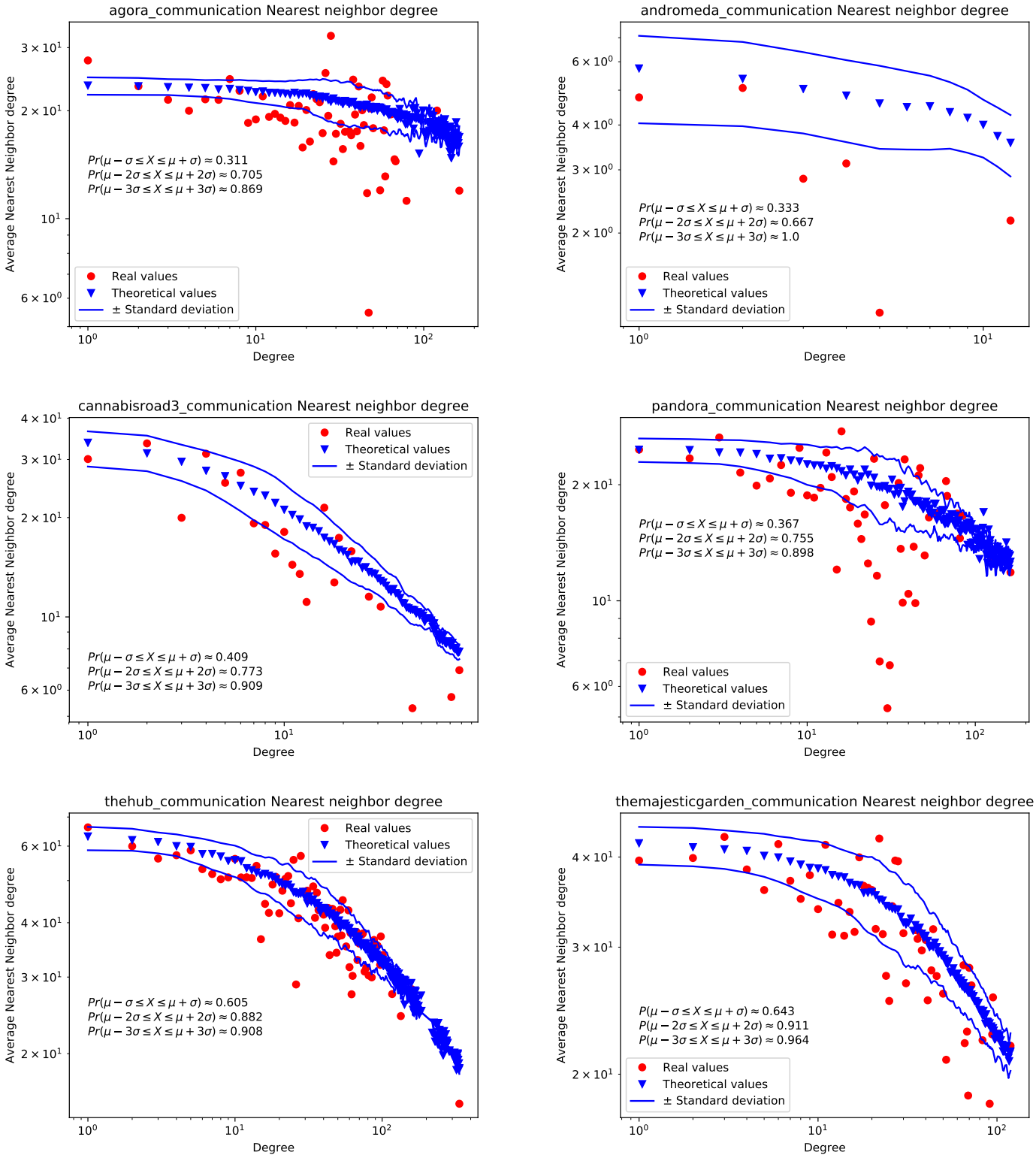

Figure 5.9: *Plots of average nearest neighbor degrees obtained by sampling*

5.2.2 Clustering coefficient

Similarly the average clustering coefficient is analysed as shown in Figure 5.10. Here it is possible to quantitatively evaluate the difference between the discrepancies amongst the nearest neighbour degree and the clustering coefficient. The Configuration Model is far worse in predicting the latter of these two topological properties. It is also now visible that the large degree nodes are more likely to fall within a standard deviation of the theoretical values. From these networks agora fairs the worst against the clustering coefficient values predicted by the Configuration Model with only 18% of the real values falling within 3 standard deviations. For this reason this network is the basis for further investigations.

5.2.3 Network growth analysis

From the results in Figures 5.5, 5.6, 5.7 and 5.8 it seems like the larger networks disagree more with the Configuration Model predictions. This leads to a hypothesis that perhaps this type of behaviour is something that becomes more apparent over time as the network grows. To test this hypothesis I used the fact in the data that I had obtained by crawling the darknet forums each event of conversation is market by its time of occurrence. This allowed me to create 10 snapshots of the network at 10 different times each separated by a period of growth by $\frac{1}{10}$ of the total number of nodes in the network. Figure 5.11 shows the real values of the average clustering coefficient per degree and the sampled ensemble averages $\pm$ one standard deviation.

The analysis shows that the responsible factors for the network topological structure are present early on - already when the network was $\frac{1}{10}$ its final size. This clearly disproves the hypothesis stated earlier.

5.2.4 Directed graphs

In the next step the directed case of the method described in Section 4.1 is used on two directed versions of networks. For the directed case I use the definition of the clustering coefficient described in Subsection 3.5.2.

Both values for the average clustering coefficient c_i^{in} and c_i^{out} are graphed as functions of the in-degree and out-degree respectively. Figure 5.12 shows these graphs for two darknet market networks: agora and kingdom.

Version of September 18, 2017– Created September 18, 2017 - 14:03

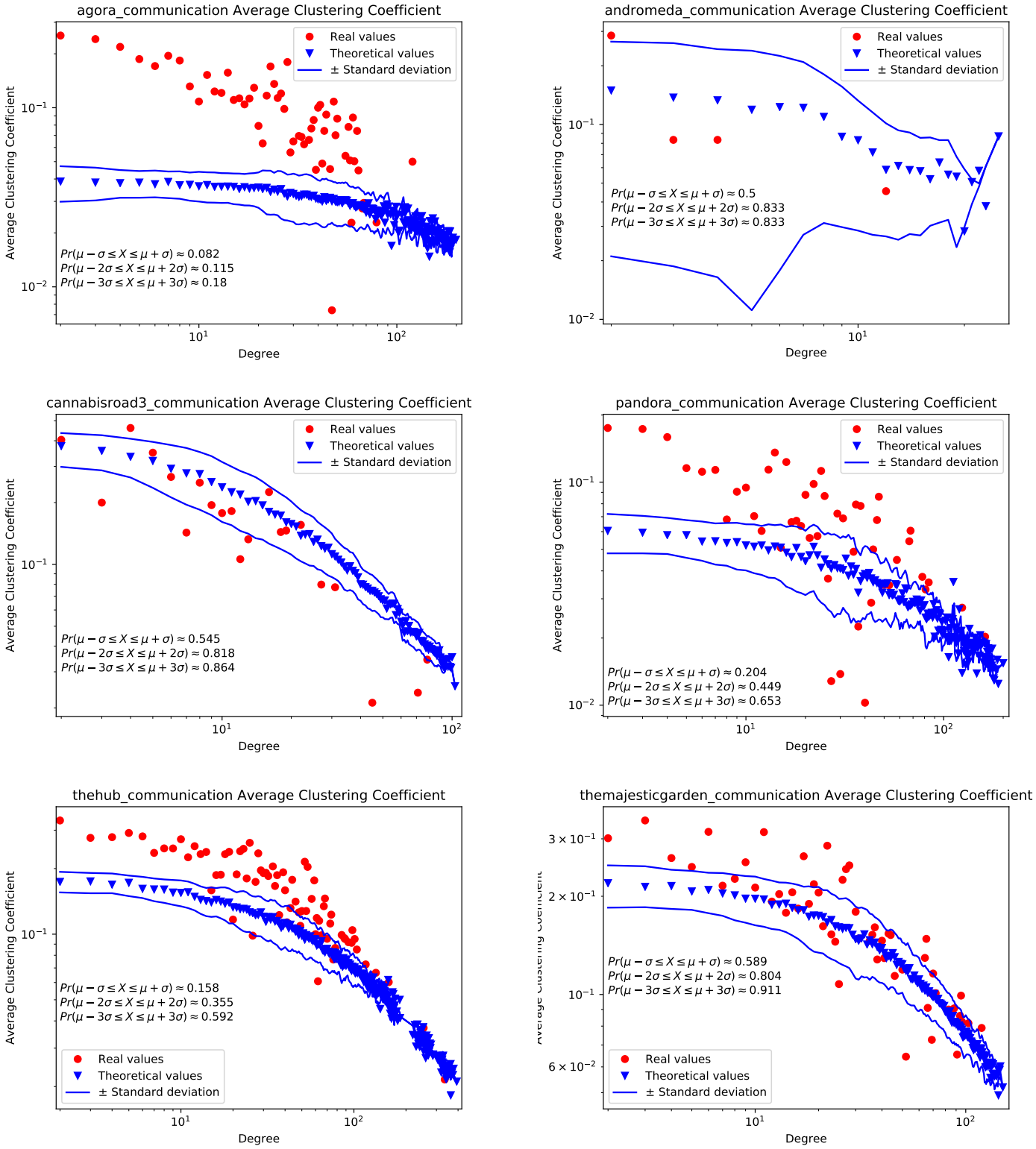

Figure 5.10: *Plots of average clustering coefficients obtained by sampling*

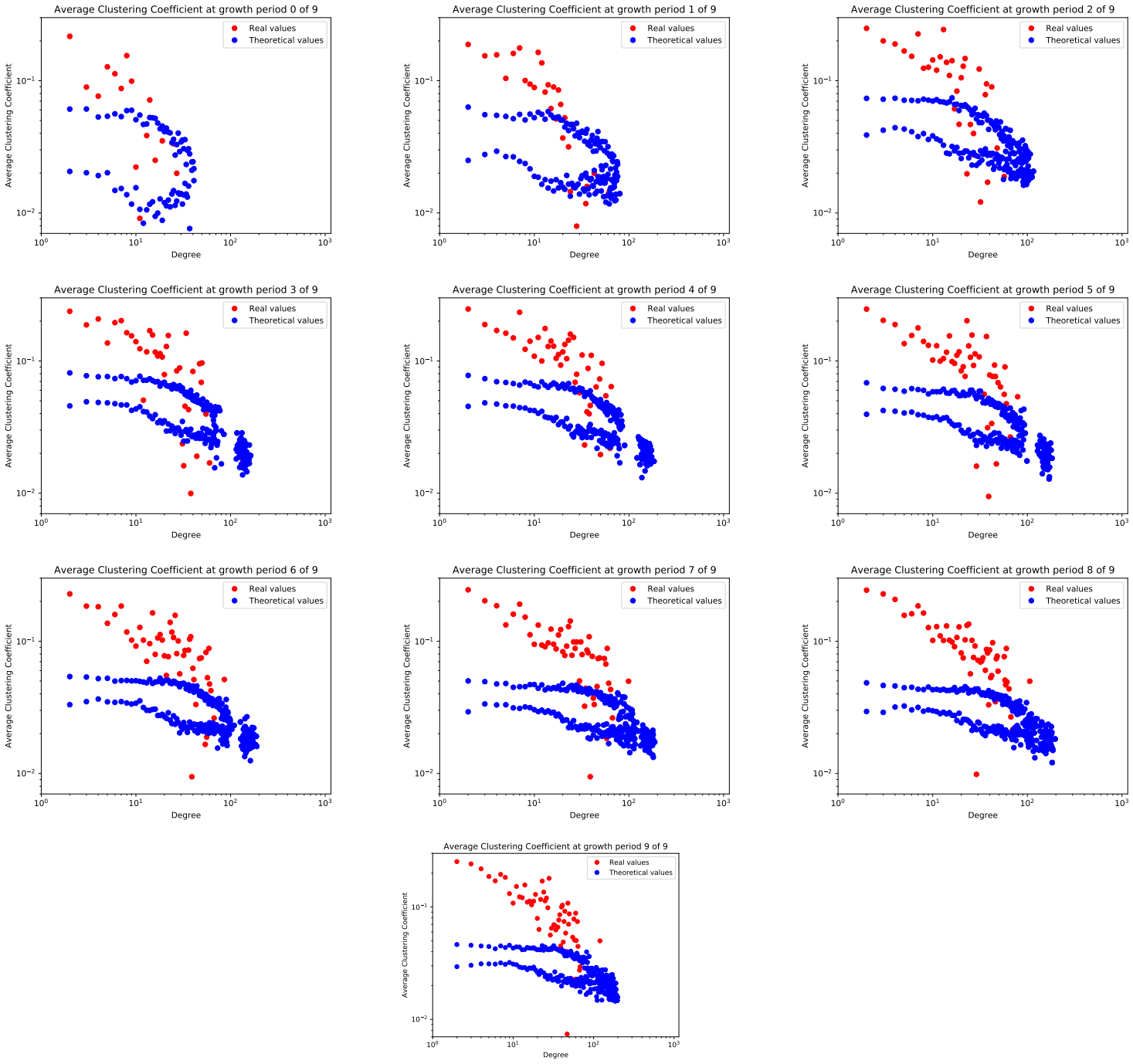

Figure 5.11: *agora average clustering coefficient as a function of node degrees over 10 periods of time*

5.2.5 Bipartite graphs

Bipartite graphs are the bipartite projections of the Expertise networks defined in Section 2.3.2. To compare them with their subsequent null model predictions the method described in Section 4.1.3 is used. For bipartite graph analysis I plot the directed in-degree versus average-out degree of nodes and the directed out-degree versus average in-degree of nodes. It is worth explaining what each of these two types of plots show.

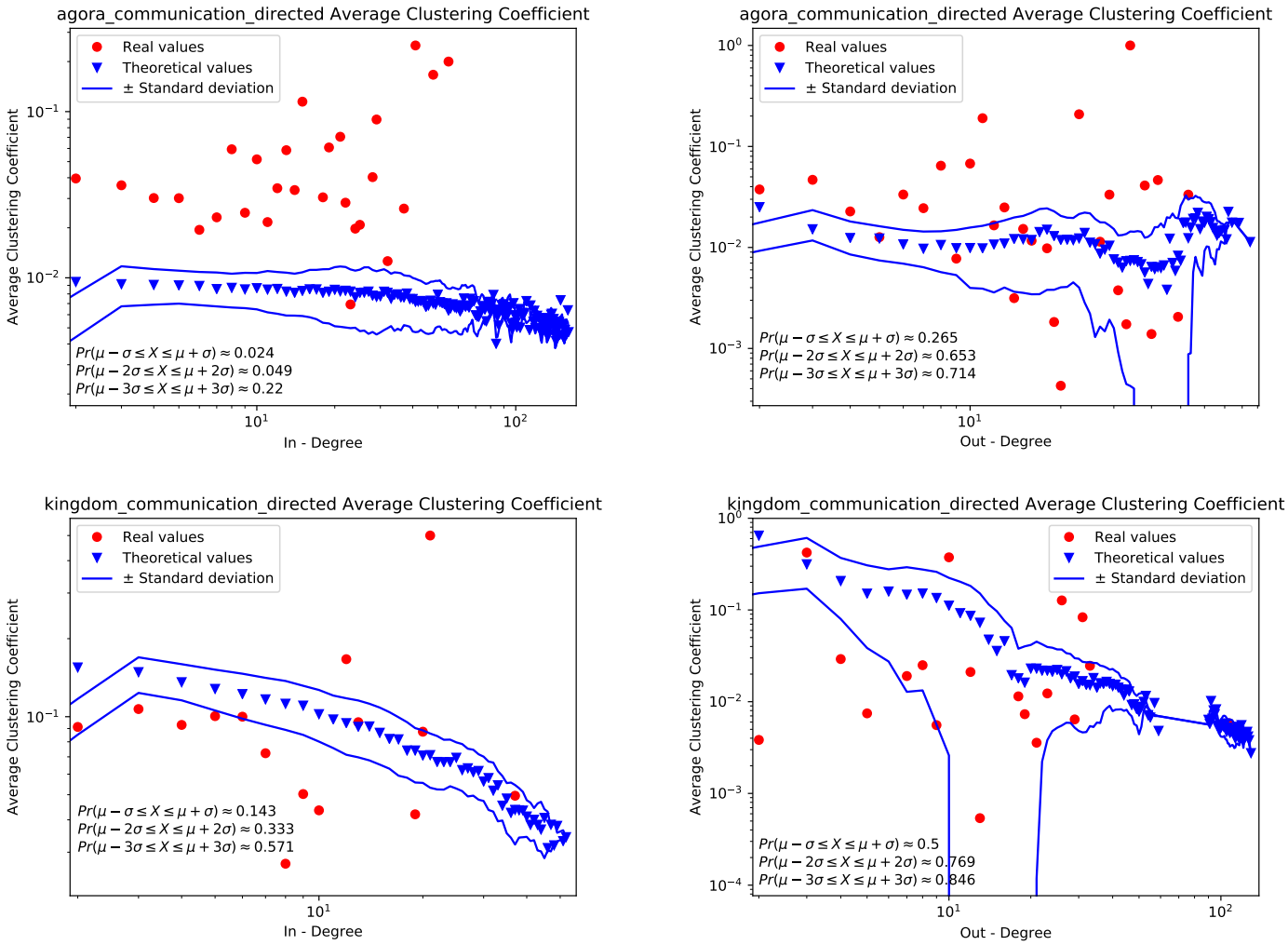

Figure 5.12: *Plots of average clustering coefficient for the directed case*

As the bipartite projection is such that all edges point to users, in other words; the user part of the networks nodes only has in-degrees and the topic part only has out-degrees. In doing the calculation and plotting the previously described topological properties we obtain two types of graphs:

Topic inclusiveness This is a graph of the out-degree of nodes in the directed projection versus the average in-degree of those out-neighbours. It shows the numbers of users in a topic and how that correlates with the average number of different topics these users participate in. Both graphs for agora and kingdom are shown on the left of Figure 5.13

User eminence These graphs show the in-degree of the same projection versus the averageout-degree. It shows the number of topics a user participates in and how that

correlates with the average number of other users that take part in these topics. These graphs are shown on the right of Figure 5.13

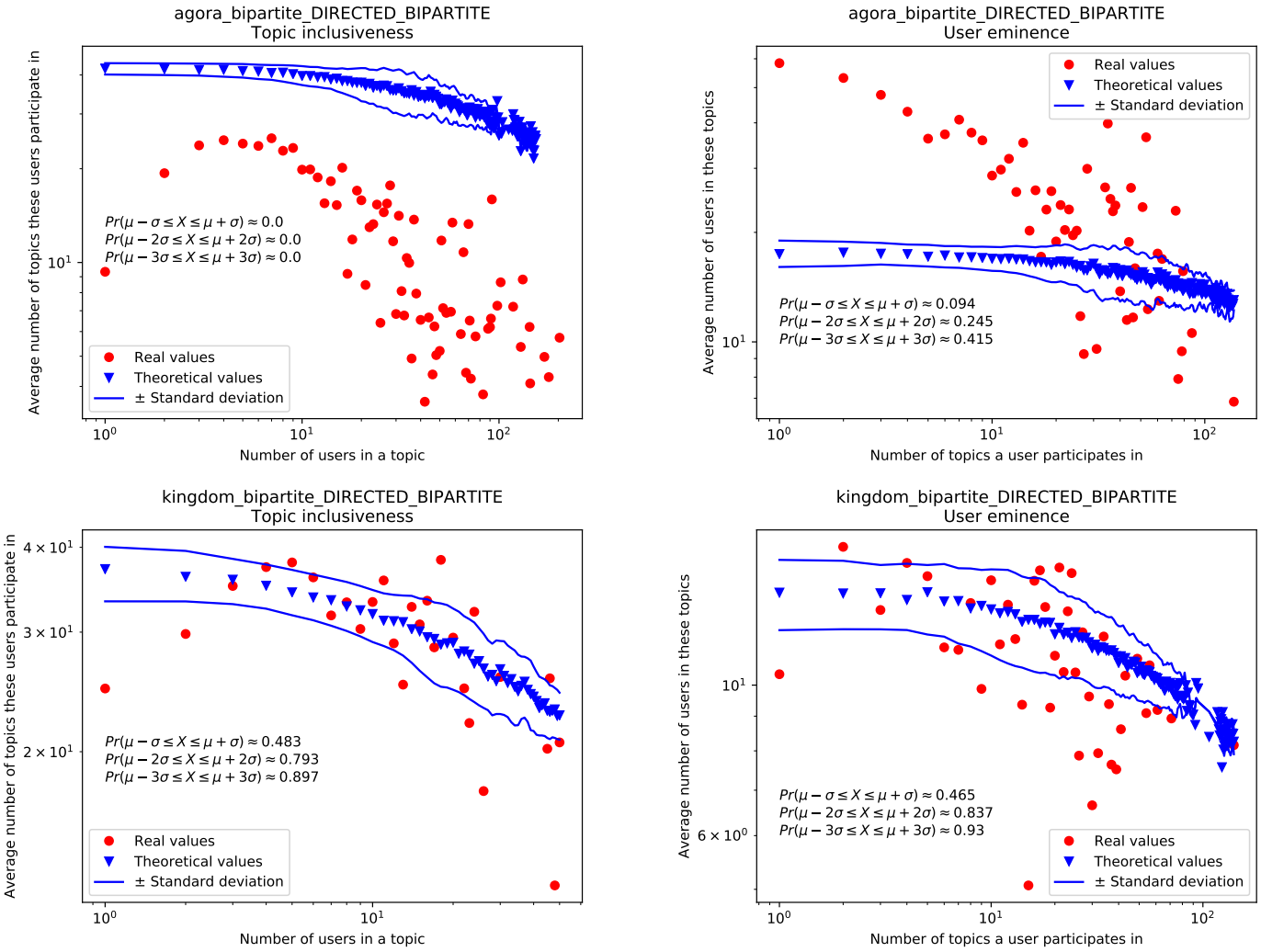

Figure 5.13: *Bipartite network comparison with the Configurtion Model predicitons*

Chapter 6

Conclusions and Discussion

The final part of this thesis discusses the obtained results and their implications. Conclusions are made with respect to the specific case of networks analysed and suitable alternatives to the configuration model as a null model are proposed and argued for.

6.1 Conclusions

The Configuration Model has shown in the past to model very well the network structures of many phenomena, amongst which are also networks of social structure. In its core lies the assumption that different actors (different nodes on the network) engage in an amount of communication (represented as edges on the graph) which follows a power-law distribution discussed in Section 3.3. But these connections themselves are inherently random. Thus for a fixed degree sequence an ensemble of uniformly random graphs should exhibit on average the same topological properties, if the network analysed can be explained by the null model.

The results I obtained in my work are interesting in two senses. For one - for simpler topological properties some of the graphs do exhibit the type of behaviour the null model predicts, so it can be as an indication that the configuration model is a good starting point in understanding the specifics of the network. However for more complex topological properties the predictions diverge, and for some - quite greatly. But it is the way how these properties diverge which can give a greater insight of the key differences between regular and criminal social networks, and some interesting conclusions of human behaviour.

6.1.1 Conclusions from the analytical approach

The analytical maximum likelihood method provided the first insights of the distinctions amongst the predicted configuration model results and the real ones. The first clue comes from the comparison of Figures 5.1, 5.2, 5.3 and 5.4 with Figures 5.5, 5.6, 5.7 and 5.8. It is clear that the more subtle and complex network structure emerges only in the third order topological property, the clustering coefficient. Up until then with the first order property of power-law degree correlations and the second order property of average nearest neighbour degrees the darknet black market networks seem to exhibit a behaviour closely resembling that which one might get assuming the configuration model as the null model. The most outstanding examples where the predicted clustering coefficient differs from the real network ones exhibit all a similar discrepancy. The real data form a steeper curve than that of the theoretical results, with nodes that have small degrees having very much higher than expected clustering coefficients. In some cases it seems that the both curves begin to converge in the high degree state suggesting that in these networks the higher degree nodes (the most 'popular' users) do behave more like the configuration model predicts they should. However the nodes with small degrees tend to cluster more than the theory predicts.

6.1.2 Conclusions from the sampling approach

The sampling approach provided an easier way to probe the dynamics of these darknet networks as well as checking the original results for errors. It helped to rapidly disprove false hypothesis, like the example described in Section 5.2.3, and to quantify the actual discrepancies from the null model with statistical analysis. This resulted in a powerful analysis technique that combined two approaches. The first approach quantified the amount of sets of equal degree nodes that fall within one, two and three standard deviations from the null model, allowing to assign values in the form of decimal fractions that explicitly present the congruity between the Configuration model and the real world networks. The second approach is the graphical representation of the average values of the topological properties of these sets superimposed on the graphs of the theoretical values. This approach was able to show that even in the cases where more than half of the data points are more than two standard deviations away from the theoretical value, they still exhibit a structure, and are not randomly distributed around the the null model predictions. This structure still seems to form a

curve, only this curve in the clustering coefficient case is a lot steeper than the theoretical one, starting with larger values for smaller degrees only to some-what converge to the theoretical curve towards the larger degrees.

6.2 Discussion

It is clear from the conclusions that to describe darknet black market network structures it is not enough to constrain only the first order properties. This means that the Configuration Model is not sufficient to describe the formation of these networks. The results of the clustering coefficient analysis reveal that in order to generate random ensembles of networks that match more closely the topology of the real networks, one must also seek to control the average number of triangles in the graph. One such model that does so and perhaps might reflect the properties of darknet black markets better is the Strauss model [6]. In this model one constrains the total number of links and total number of triangles (node triplets that are all connected to one-another) in the graph.

The fact that smaller degree nodes tend to cluster more than the Configuration Model predicts and the convergence to this null model in the higher degree case might offer some insights in the interpretation of this behaviour. Because these are notorious networks in which most of the actions are illegal, occasional actors will try to limit their communication to not be associated with to much illicit behaviour. So those that have a few questions or comments about a certain topic will tend to stay within that topic, and thus tighter communities will form amongst these small degree actors. This is also visible in the bipartite user-topic network analysis. The user eminence has a steeper correlation than the original null model predicts, showing that these users participate in far less topics than predicted. However, when we look at high degree nodes - users with a lot of communication, the graphs of the clustering coefficient tend to converge to what the model predicts. This shows that regular users tend to behave in a way that resembles usual social network behaviour. It would be reasonable to assume that these users are those who are actually selling things on these marketplaces, or are the owners of the marketplace themselves, either way those who are profiting off of this illicit business.

To sum up I conclude that darknet market users who profit from participating in these markets (and by doing so are the most active users) tend to behave as classic social network users, but those who are more passive users, more likely to just purchase

illicit goods from time to time form denser clusters in their communication.

Bibliography

[1] Pharmaceutical crime on the darknet a study of illicit online marketplaces. INTERPOL Criminal Analysis Sub-directorate report, February 2015.

[2] Albert-Laszlo Barabasi. *Network Science*. Cambridge University Press, 2015.

[3] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 1999.

[4] Monica J. Barratt, Jason A. Ferris, and Adam R. Winstock. Use of silk road, the online drug marketplace, in the united kingdom, australia and the united states. *Addiction*, 2013.

[5] Francesco Buccafurri, Gianluca Lax, and Antonino Nocera. A new form of assortativity in online social networks. *ELSEVIER*, 2015.

[6] Ton Coolen, Alessia Annibale, and Ekaterina Roberts. *Generating random networks and graphs*. Oxford university press, 2017.

[7] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[8] Brewster Kahle. Darknet Black Markets. https://archive.org/download/dnmarchives, 2016. [Online; accessed 19-July-2016].

[9] Silvio Lattanzi and Yaron Singer. The power of random neighbors in social networks. *Proceeding WSDM '15 Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015.

[10] James Martin. *Drugs on the Dark Net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs*. Palgrave Macmillan UK, 2014.

[11] Peter R. Monge and Noshir S. Contractor. *Theories of Communication Networks*. Oxford university press, 2003.

[12] M. E. J. Newman. Assortative mixing in networks. *Physical Review*, 2002.

[13] M. E. J. Newman. Mixing patterns in networks. *Physical Review*, 2003.

[14] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review*, 2003.

[15] Tiziano Squartini and Diego Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 2011.

[16] Tiziano Squartini, Rossana Mastrandrea, and Diego Garlaschelli. Unbiased sampling of network ensembles. *New Journal of Physics*, 2015.

[17] D. J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.

[18] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: Structure and algorithms. *Proceeding WWW '07 Proceedings of the 16th international conference on World Wide Web*, 2007.