AMSI
AUSTRALIAN MATHEMATICAL
SCIENCES INSTITUTE

National
Collaboration
in the Mathematical
Sciences

The Statistical Properties of Complex Networks

Michael Mcphail
Supervisor: Professor Michael Small

June 22, 2014

1 Introduction

When working with large real world systems such as the wiring of a brain or the internet, the amount of data contained in such a system can make obtaining useful information difficult. By representing a large and complicated real world system in a much simpler form of points (vertices/nodes) connected by lines (edges) which we call a network, we can often obtain information on the system as a whole. Complex networks are large heterogeneous networks that have features not typically found in random (erd ös-rényi networks). Many real world networks are categorised as complex networks because they contain the kind of features that random networks typically don't contain [5]. In order to understand more about real world networks it would be useful to be able to compare them to similar networks with some similar constraints. The current method of generating networks makes direct comparison not so straight forward because it is not often clear whether we are generating suitable networks. The purpose of this report is to investigate a new method of network generation and comparison which directly conserves one of the initial properties of a network in question. In order to test this network generation algorithm, some real world networks have been selected and a select number of network properties have been used as initial test properties to find out what this algorithm can tell us. This test will either show that a real world networks' measured properties fall within a statistically acceptable range from the distribution of the values measured on our generated networks, or that certain properties of a real world network are outliers of such a distribution. In the second case an interesting feature of our real world network may have been discovered.

2 Terminology

Node / Vertex -Points that are used to represent an individual, group or object. In all of the cases studied below nodes have only one type, for example in an airline network all nodes represent airports or in a social network all nodes may represent people.
Edge / link - If an edge exists between two nodes then this symbolises some relationship exists between these two nodes. For example edges are used to represent flights between airports in an airport network.
Network - A network is a collection of nodes and their associated edges, networks are often used to represent a real world system
Degree of a node - The number of edges attached to a node
Degree distribution - The number of nodes in a network of each degree.

3 Background

One particular class of networks currently studied is "scale free" networks which are categorised by having degree distributions that follow a power law. A lot of real world complex networks fall into this category of being scale free. The current method of generating scale free networks is called preferential attachment which involves adding nodes to a network along with a fixed number of edges attached to this node, the probability that one of these edges attaches to one of the networks current nodes is proportional the current nodes degree. Higher degree nodes have a higher probability that a new edge will be attached to it and therefore the degree of these nodes increases faster than the nodes with smaller degree. This produces a degree distribution with a high number of nodes with small degree and a non zero probability of there being a node of arbitrary degree. This method of network generation can generate all networks that result from a given growth mechanism, which forms the sample space if we were to sample from networks generated this way. While using this method it is reasonably easy to generate scale free networks of arbitrary size and by tweaking the number of edges you add each time or possibly adding constraints, the growth mechanism sampled from can be changed. However with real world networks it is often difficult to determine how a network has grown so that even your best estimate of a networks growth mechanism could still give you a sample space that is very different from an experimental network. For direct comparison it would be useful to have a set of networks with which it is known share similar constraints and have some constraints (that we may not be able to quantify in real networks) left free.

One such network property that can be conserved is the degree distribution of the network. Instead of sampling from all networks of a given growth mechanism, we sample from all networks of a given degree distribution (say the degree distribution of the initial real world network). This means that we can make direct comparison to networks which share at least one property and see if our experimental network is typical of such a distribution. The properties of the networks in this sample space would be expected to follow a probability distribution and it is possible that we may find that real world networks have properties that are outliers from from this distribution. If such a property were to be found there may be an underlying reason as to why this occurs such as an constraints on the system or a particular purpose a network is designed to fulfil. So finding such properties allows for the opportunity to discover and investigate special properties of real world networks.

4 Method

In order to randomly generate a network with the same degree distribution as the real world network but with independent structure, an algorithm is used whose function is to switch edges between unconnected pairs of connected vertices (see Figure 1 below).

This process changes the local structure of the network but not the degree distribution of the network as all nodes are still attached to the same number of edges. By repeatedly applying this process the degree distribution remains the same, but eventually the wiring of the network becomes independent of the initial structure of the real world network. In order to determine approximatelly when the generated network becomes independent of the initial network, the edge-switch algorithm can be run repeatedly while periodcally taking measurements of the properties that we are interested in. What is observed is that the properties start at some initial value and then tend towards some distribution. This typically looks like the plot in Figure 2 below.

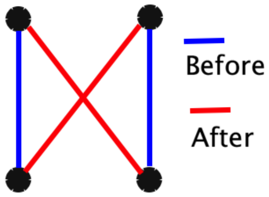


Figure 1: shows the wiring between the four nodes involved in an edge switch before and after the process

The flat region in this plot indicates that the algorithm is now sampling from networks according to some equillibrium distribution implying the structure is independent of the initial networks wiring. Once an independent network is generated its properties can then be measured and if this process is repeated many times than the measurements of these distributions forms a probability distribution. Another very similar but computationally less intensive method is to start with the initial real world network and obtain the first independent network, then perform the edge switch on this network until a new network is obtained that is independent of the previously generated network. This usually doesn't take as long to generate multiple independent networks because the network already starts within the required equilibrium distribution.

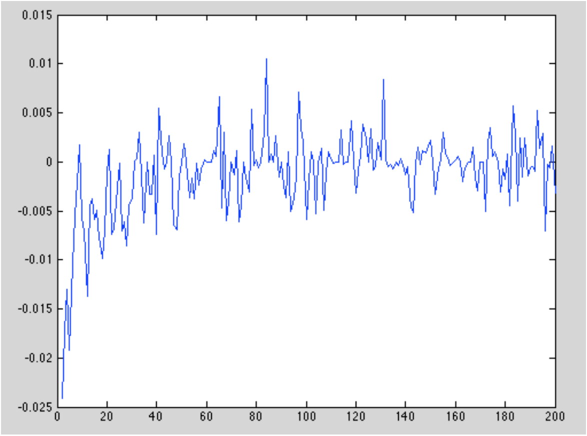


Figure 2: shows the difference between a measurement of the clustering of a network before and after approximatelly 3000 edge switches. This process was repeated 200 times and the x-axis indicates measurement number.

By altering our initial network within the given state space of all networks with a given degree distribution until we reach an equilibrium distribution, effectively a sample space of all possible networks with the given initial degree distribution has been built (essentially a Markov chain Monte Carlo [6]). By sampling 1000 random realisations of this sample space a distribution of the properties of these network forms. Then the distribution obtained can be compared to the properties of the original network.

In order to properly analyse a real network, network properties that represent the physical structure of a network should be measured and compared in the real world system and in the generated system. The properties that will be compared in the rest of this paper are:

1. Clustering:
   Roughly equates to the probability that the neighbours of a common node are themselves neighbours. This property is calculated by finding the clustering co-

efficient (the ratio of connected paths of length three to total possible connected paths of length three) at each vertex and then calculating the mean for all vertices. In a social network this would be equal to the probability that two of an individuals friends are friends with each other as well.

2. Assortativity:
   The likelihood that nodes of similar degree are connected. It is the correlation between the degree of the nodes at either end of an edge, assortativity close to 0 indicates no definite correlation between the degrees of connected nodes. If assortativity is greater than 0 the network is assortative and if it is less than 0 it is disassortative. This property takes a value between -1 and 1 with the extreme values indicating a strong correlation between the nodes degrees and there likelihoods to be connected to each other. For example, how likely are two nodes of high degree to be connected.

3. Diameter:
   The median of the mean of the minimum path length between random nodes. Essentially used as a measure of distance.

The reason that these properties were chosen as opposed to other possible properties is because they are properties of the network as a whole and not just of the individual components. This allows for analysis into how the whole network changes and not just the structure around some particular component or node.

This technique of network generation is useful in that it can be used to analyse real world networks. In order to do this, some real world networks that represent some of the main classes of systems have been selected for testing. These networks include;

1. Airline network - A network containing information on the 500 busiest airlines in the United States, in this network the nodes represent airports and the edges represent flights between airports. This is an example of a technological network. [1]

2. Neural network - A network representing the neural structure of a nematode called C.elegans (Caenorhabditis elegans) in which edges represent links between neurons. This is an example of a biological network. [2]

3. High school network - A network demonstrating the friendships between school students in a rural American high school. Nodes represent students and edges represent friendships between students. This network is an example of a social network.[3]

4. Co-authorship network in network science - A network that demonstrates co-athorship of papers between network scientists. Nodes represent network scientists and edges represent that these two scientists have coauthored a paper together. This network is also a social network.[4]

The large complicated nature of complex networks make obtaining any information by simple observation difficult. The networks mentioned above can be visualised without any constraints encoded using Mathematica's graphing capabilities (see appendix 1) The degree distributions of these networks (see appendix 2) form the constraints for the networks that we sample from to make comparisons with our real world network.

5 Results

After using the previously described method on the Airline network, the C.elegans neural network, the high school network and the co- authorship network, distributions were obtained and compared to the real world measurements (see appendices 3, 4, 5 and 6 respectively for visualisations of these results). What is clear is that there is some very evident "outliers", properties that do not seem typical of the distributions of values. The presence of such an outlier indicates that the structure of the network is special in some way and ideally it would be helpful to consider physical explanations as to why it is special.

5.1 Airline Network

The properties of this network that are not typical of networks of the same degree distribution are clustering and diameter. This implies that there is some kind of constraint or other factor that drives network to grow in a very atypical way. The clustering of the real world airline network is much greater than any of the values measured from the random realisations of its degree distribution. If this situation is thought of physically it can be hypothesised that a majority of this can be attributed to the geographic constraints on the system (including distance) which makes flying between closer airports easier. An interesting feature in the comparison between the real networks diameter and that of the random realisations is that the real diameter is higher than the distribution. This is interesting because the airline network is designed by humans to transport people (not goods in this case) so a higher number of flights between random destinations would be undesirable. This means that either the current flight path design is inefficient or that other constraints on the system make it so that more flights must be taken than average given the degree distribution of the network.

5.2 C.elegans neural Network

The results of this network show that all measured properties were outliers. We could hypothesise in this case that the higher clustering is a result of the way the nematodes brain must process information (i.e. its purpose). The assortativity in this case, while negative, is still close to zero meaning there is still not a large amount of correlation between the degree of a node and the degree of another node connected to this node. The comparison between the random realisations and the real measure value of diameter is interesting in that while it shows the real value to be atypical, the difference between the mean value of the distribution and the measured value is 0.06. In a physical sense this is not likely to have a noticeable impact on the structure of the nematodes neural network. This shows that even though a property is an outlier to the distribution of measured values, this may not always have a large impact on the physical functions of a system.

5.3 Highschool Network

The high school network is a social network which means it is expected that such a network should have a large clustering value. This usually arises because it is more likely that an individuals friends are themselves friends (due to a common contact and more likelihood of interaction e.c.t.). What is shown by the results of the high school network analysis is that the clustering is in fact much higher than that of the random realisations. There is no definition of what a high clustering value actually is, if it is possible to have some similar networks to compare to, it may be easier to determine whether or not the clustering of a network is actually higher than it needs to be. It is also noticeable that the diameter of this network is quite high, possibly due to the large amount of clustering that occurs.

5.4 Co-authorship Network

The co-authorship network is again a social network, so it would be expected to be in some ways similar to the high school network. What is immediately obvious when comparing the results for the two networks is that in both cases both the diameters and the clustering values were much higher than the values measured for the random realisations of their respective degree distributions. This may be a consistent property of all social networks and provide a reliable way to compare and categorise social networks.

This analysis produced a lot of outliers when comparing the properties of a real world network to the properties of the random realisations of that networks degree distribution. This suggests that real world networks seem to be highly atypical and it was found that some are more atypical than others. For example it seems that the properties measured on the co-authorship network were much further from the distribution of the measured values than the neural network seemed to be. One reason that such a large proportion of the properties were outliers could be because the purposes and functions of real world networks make them specialised in a range of categories. It is possible that this method is more suited to specific classes of complex networks such as the airline network which was a technological network. Not only did the analysis of this network provide an interesting result related to the diameter, physically performing an edge switch on this network simply involves changing future flight plans (i.e. it is physically plausible to do this). Another possible explanation as to why so many of the network properties were atypical could be the fact that these properties were selected, and they may have been unintentionally selected on that expectation that these properties were what made the system in question special. It would be useful to measure a lot more properties, some which may not be completely physically significant, so as to determine whether real networks are atypical in a wider range of properties.

6 Conclusion

The method of switching the edges of real networks to construct a sample space of all networks with the same degree distribution as the real network has the possibility to confirm expectations about some networks as well as uncovering previously unexpected properties of others. An example of a property of a network that may have been unexpected was the airline network analysed which had a diameter higher than any of the random realisations measured. The reason this property may have been unexpected is because amongst the likely constraints on the system there would be geographic ones as well as those enforced by humans as to build an efficient system. Both of those constraints would have been conflicting and to be able to recognise that similar networks have lower diameters indicates that the effect of the geographic constraints are stronger. Another interesting result of this technique was the confirmation of the similarities between networks of similar class. It was found that both the high school network and the co-authorship network in network science had clustering and diameters far above most other networks with the same degree distribution. Uncovering such similarities between networks of similar class may help to categorise unknown networks or increase our understanding of the networks in this class that we already know. What is clear from this analysis is that real world networks have properties that are not at all typical of the random realisations of the networks with the same degree distribution. This suggests that if a system has a purpose then we can often find special properties of the network which result from the physical function of the system.

7 Acknowledgements

Thank you to Professor Michael Small for offering me this project as well as providing continued assistance and guidance throughout. Thank you also to AMSI for the opportunity to work on a project such as this and for providing the opportunity to attend the Big Day In at UNSW.

8 References

[1] Colizza, V., Pastor-Satorras, R., Vespignani, A., 2007. Reaction-diffusion processes and metapopulation models in heterogeneous networks. Nature Physics 3, 276-282 (Available from https://sites.google.com/site/cxnets/usairtransportationnetwork)

[2] D. J. Watts and S. H. Strogatz, Nature 393,440-442 1998. Original experimental data taken from J. G. White, E. Southgate, J. N. Thompson, and S. Brenner,Phil. Trans. R. Soc. London 314, 1-340 (1986). (Available from http://www-personal.umich.edu/mejn/ netdata/)

[3] M.D. Resnick, P.S. Bearman, R.W. Blum et al. (1997). Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health, Journal of the American Medical Association, 278: 823-32. (Available from http://svitsrv25.epfl.ch/R-doc/library/ergm/html/faux.mesa.high.html)

[4] M. E. J. Newman, Phys. Rev. E 74, 036104 (2006). (Available from http://www-personal.umich.edu/ mejn/netdata/)

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics. PhysicsReports 424(2006)175308

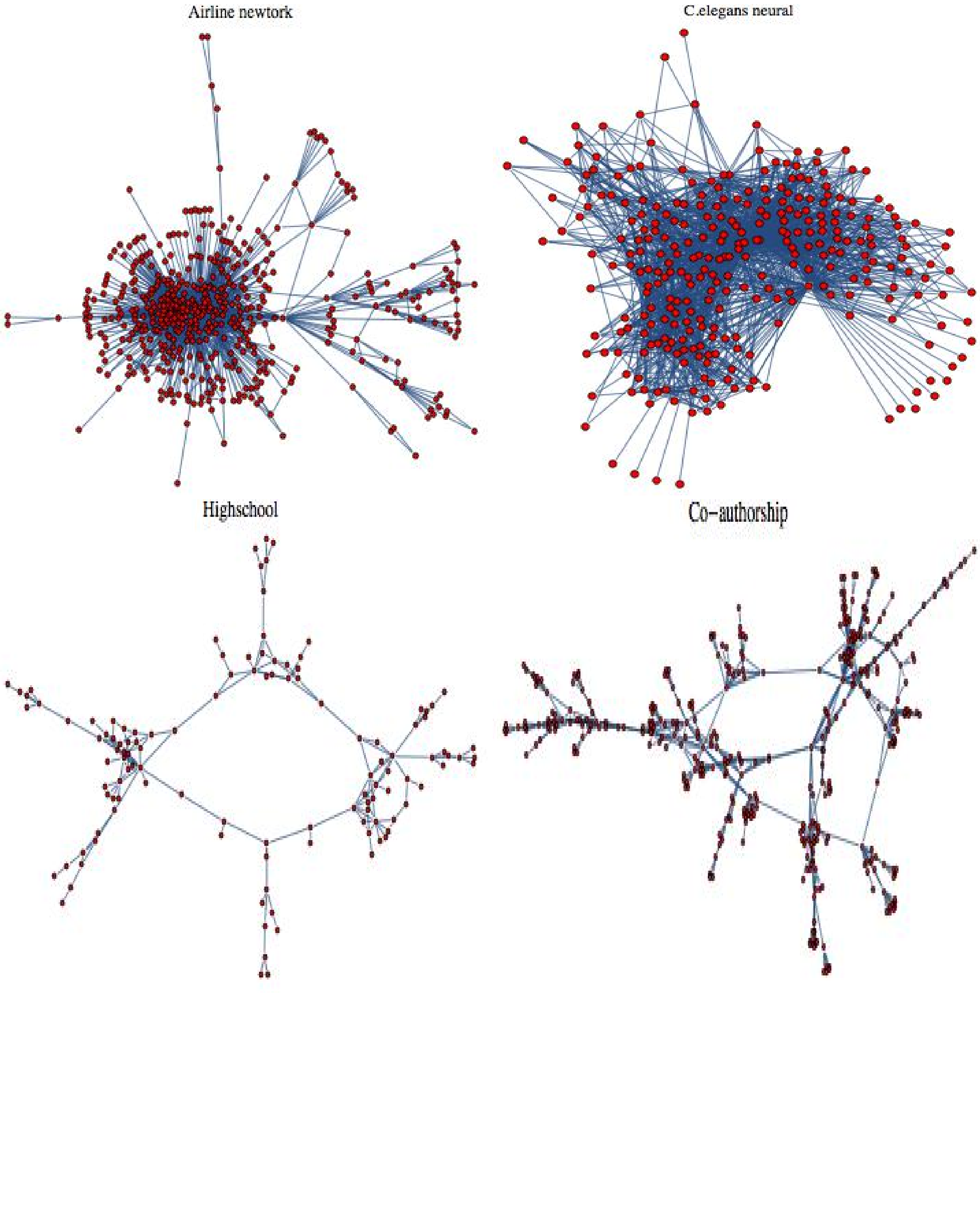
[6] Michael Small, Kevin Judd and Thomas Stemler, A surrogate for networks How scale-free is my scale-free network?, arXiv:1306.4064, (Avaibable from http://arxiv.org /abs/1306.4064)

[7] Linjun Zhang, Michael Small, and Kevin Judd, Exactly scale-free scale-free networks, arXiv:1309.0961, (Available from http://arxiv.org/abs/1309.0961)

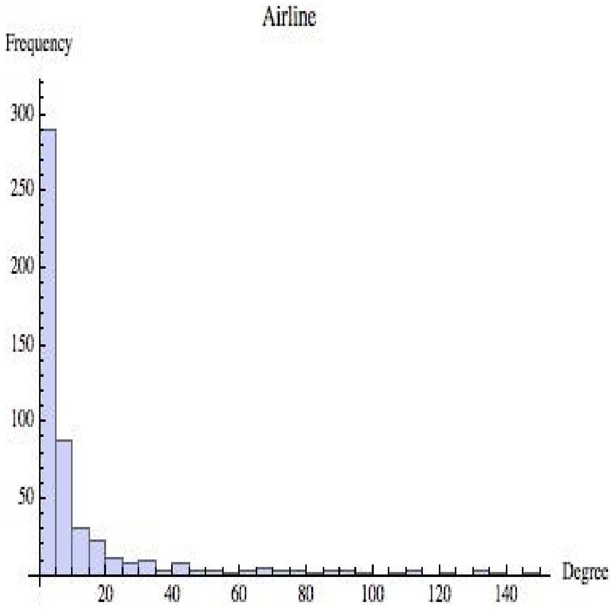
[8] M. E. J. Newman, The Structure and Function of Complex Networks, SIAM Review 45, 167-256 (2003).

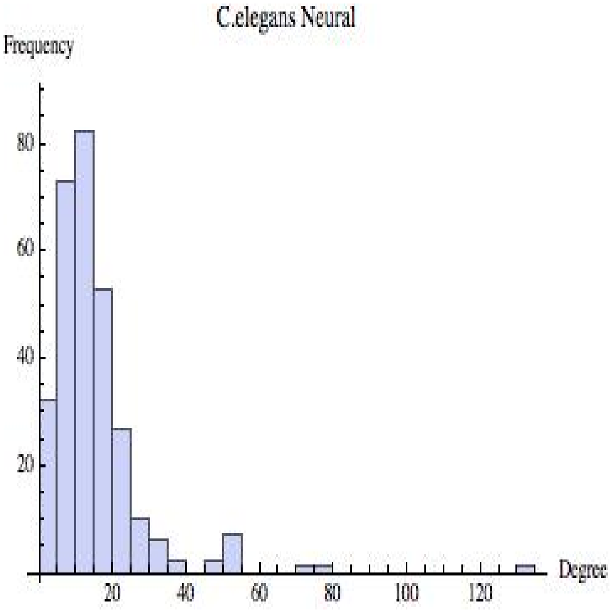
9 Appendices

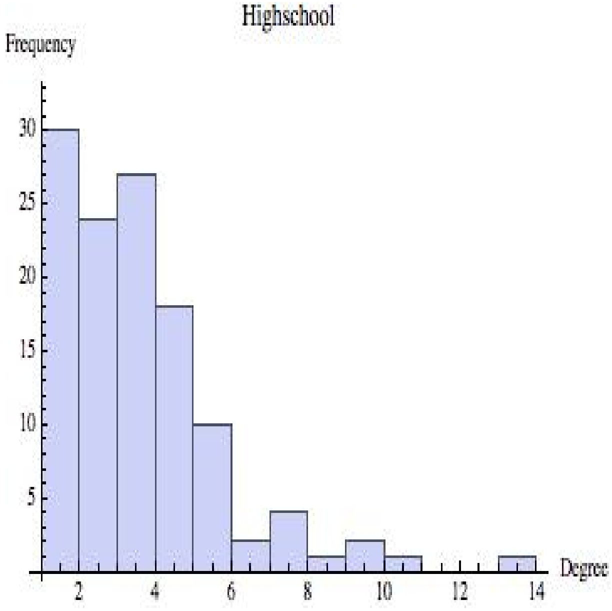
9.1 Appendix 1 - Real world networks

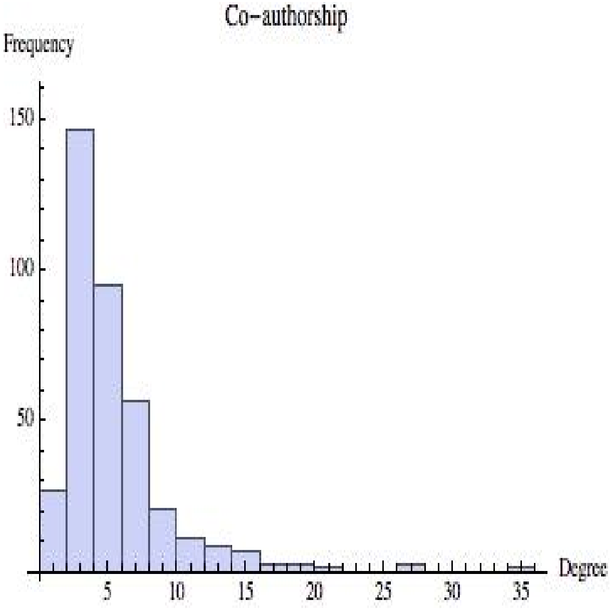


9.2 Appendix 2 - Network Degree Distributions

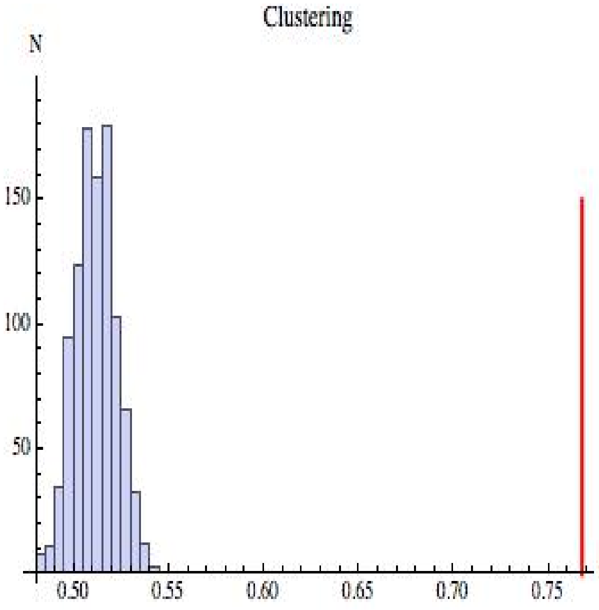

Airline

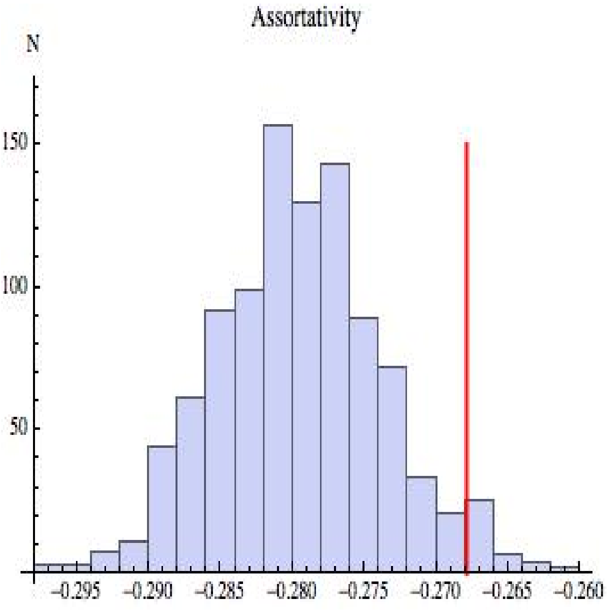

C.elegans Neural

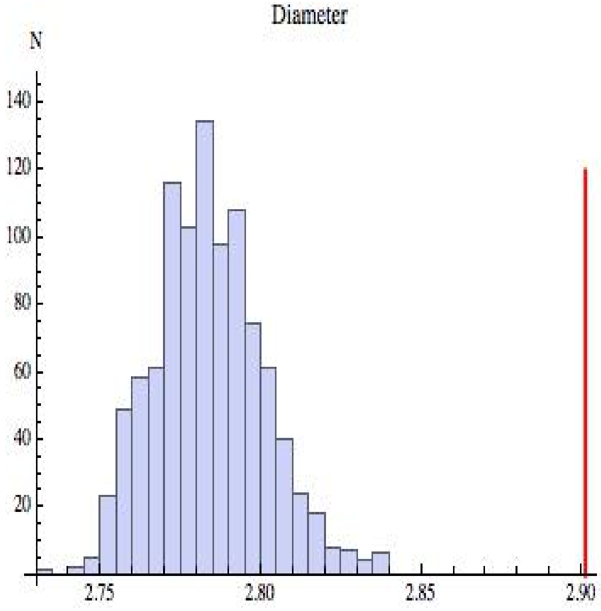

Highschool

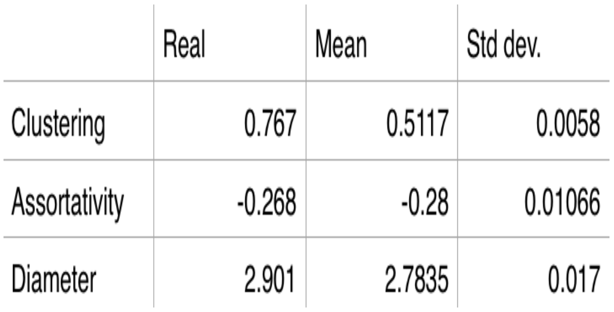

Co–authorship

9.3 Appendix 3- Airline Network





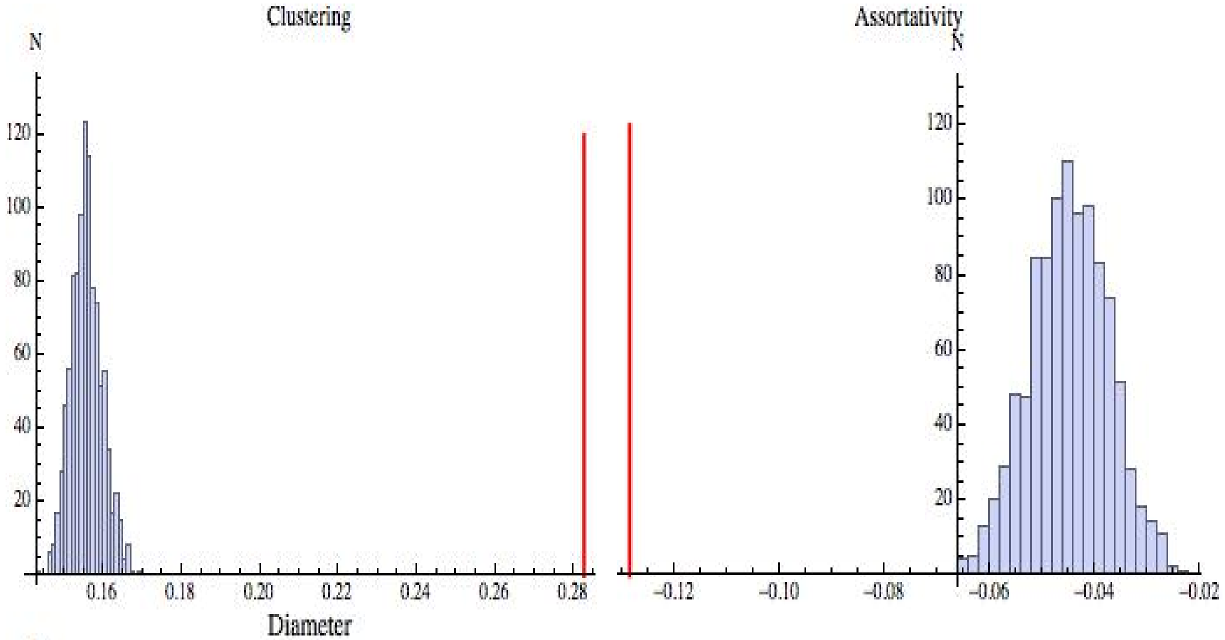


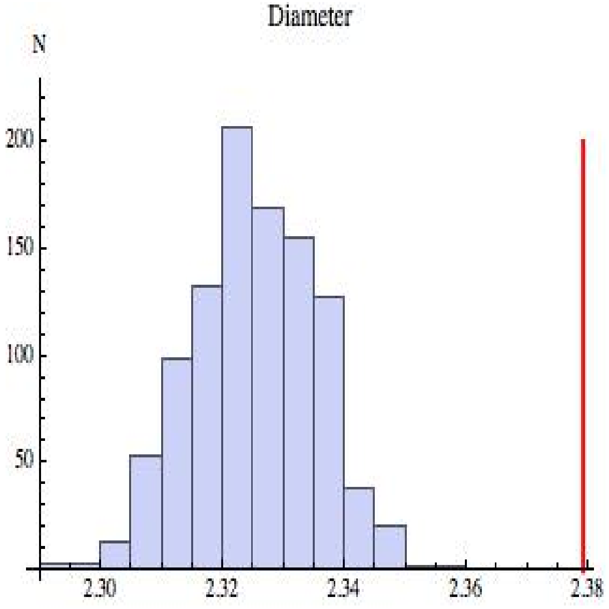


	Real	Mean	Std dev.
Clustering	0.767	0.5117	0.0058
Assortativity	-0.268	-0.28	0.01066
Diameter	2.901	2.7835	0.017

Note: "Real" indicates the measurement on the real world network, "Mean" and "Std dev." refer to the distributions of the measured values. On the histograms shown, the position of "Real" is indicated by the red line.

9.4 Appendix 4 - Neural Network

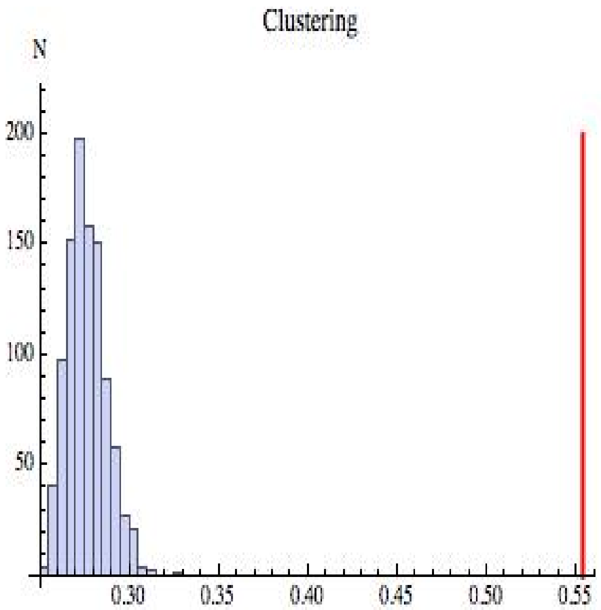


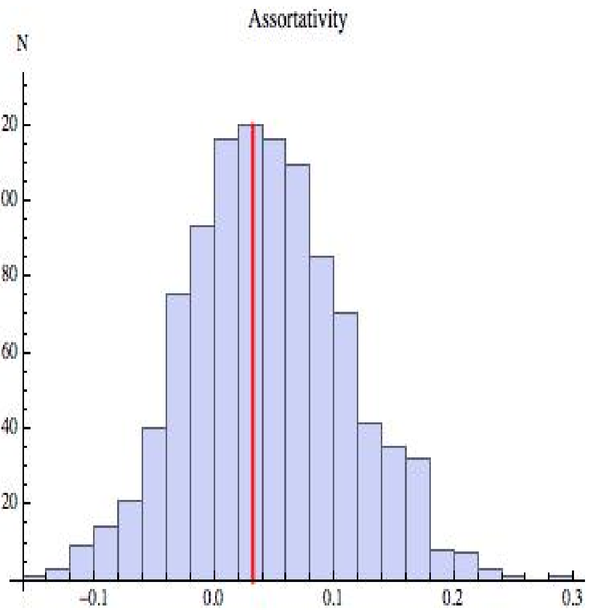


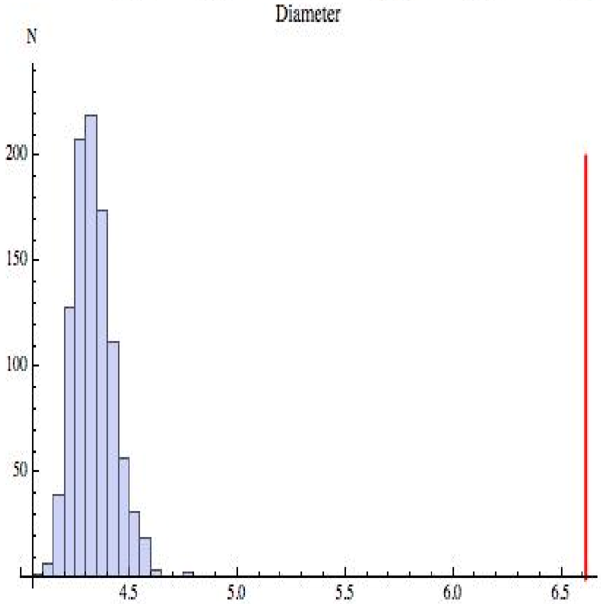
	Real	Mean	Std dev.
Clustering	0.282	-0.044	0.0075
Assortativity	-0.129	0.16	0.0039
Diameter	2.38	2.32	0.012

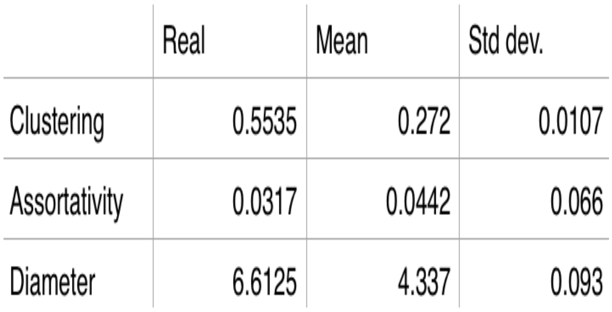
Note: “Real” indicates the measurement on the real world network, “Mean” and “Std dev.” refer to the distributions of the measured values. On the histograms shown, the position of “Real” is indicated by the red line.

9.5 Appendix 5 - High School Network





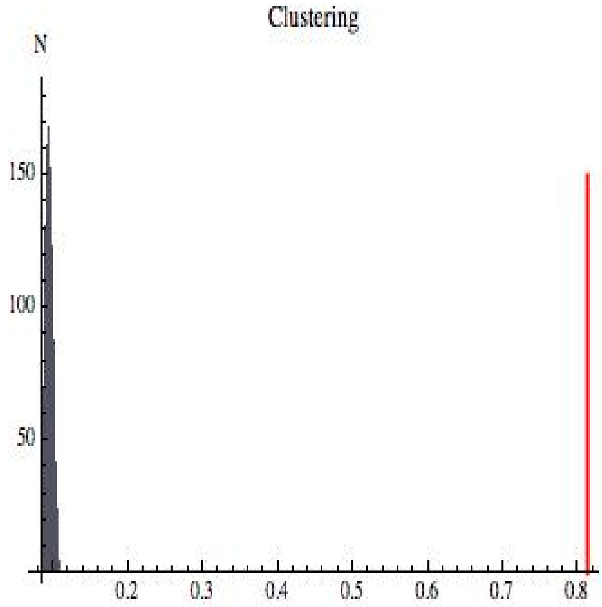


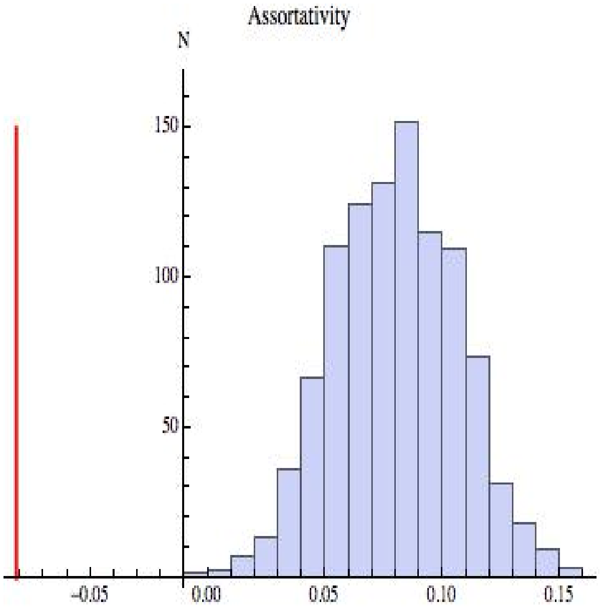


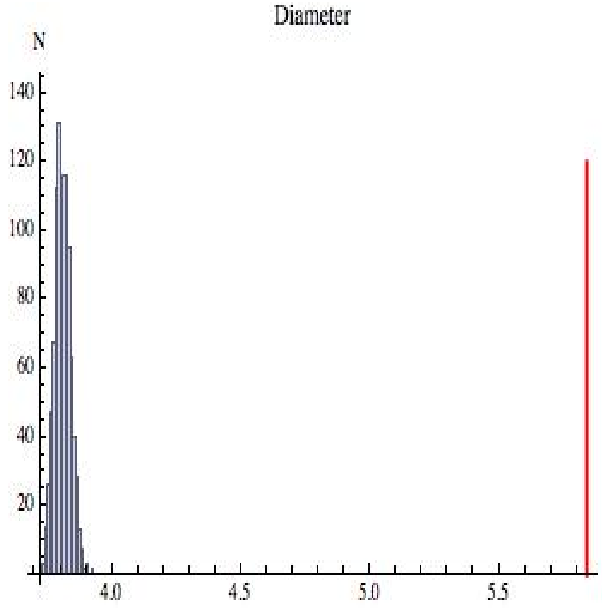
	Real	Mean	Std dev.
Clustering	0.5535	0.272	0.0107
Assortativity	0.0317	0.0442	0.066
Diameter	6.6125	4.337	0.093

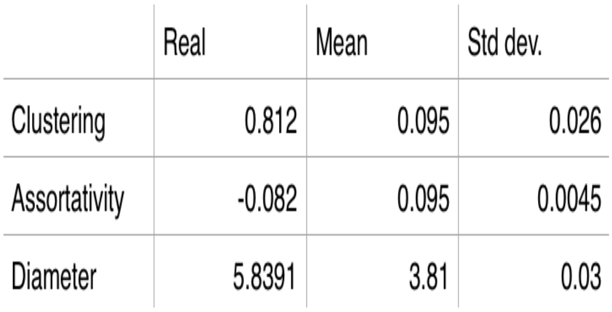
Note: "Real" indicates the measurement on the real world network, "Mean" and "Std dev." refer to the distributions of the measured values. On the histograms shown, the position of "Real" is indicated by the red line.

9.6 Appendix 6 -Coauthorship Network









	Real	Mean	Std dev.
Clustering	0.812	0.095	0.026
Assortativity	-0.082	0.095	0.0045
Diameter	5.8391	3.81	0.03

Note: “Real” indicates the measurement on the real world network, “Mean” and “Std dev.” refer to the distributions of the measured values. On the histograms shown, the position of “Real” is indicated by the red line.