2 Statistical Distribution of Networks

A major challenge in describing change in a network is to first understand the statistical distributions that describe the typical behavior of a network in the first place. This challenge has been a rich area for research for more than 60 years. For more information on network statistics the reader is referred to Wasserman and Faust (1994), Carrington, Scott, and Wasserman (2007), Barabasi, Watts, and Newman (2005). This chapter will focus on a single property of a network, the degree. The degree, k of a node i is the number of other nodes linked to node i.

A popular topological property of networks is the distribution of the degree of nodes in the network. Many have fit power law distributions to this property (Barabasi and Albert, 1999; Goh, Kahng, and Kim, 2001; Barabasi, 2002; Barabasi et al, 2002; Bollobas and Riordan, 2003; Pastor-Satorras and Vespigani, 2004; Newman, 2005; Caldarelli, 2007). This has given rise to the Scale-Free network inspired by statistical mechanics, which is characterized by having a power law degree distribution which tends to create hub nodes that are highly connected, while most nodes have relatively few connections in the network. Some have criticized this approach for social networks on the grounds that it does not consider the context of the network data and assumes universal network behavior (Doyle et al, 2005; Wasserman, Scott and Carrington, 2007; Alderson, 2008).

Another popular model for degree distribution is the random graph (Erdos and Renyi, 1959, 1960, 1961; Wasserman and Faust, 1994). Under this model, nodes in the network are connected with equal probability. This has been used in many cases as a null hypothesis to test the "randomness" of empirical network data (Wasserman and Faust, 1994; McCulloh et al, 2007; Alderson, 2008). Under this model of the distribution of node degree, the degree tends to appear normally distributed as the number of nodes increases and probability of connection remains the same. Some argue that this model is therefore not representative of empirical data (Watts and Strogatz, 1998; Barabasi, 2002; Pastor-Satorras and Vesppignani, 2004; Newman, 2005; Caldarelli, 2007).

Between these extremes, more realistic network models have been proposed by introducing more detailed models of the probability of node connection. Some of these models have been based on established social theory and multi-agent simulation (Carley, 1990, 1995, 1999; Doreian and Stokman, 1997), while others have been based on structural properties within the network (Frank, 1991; Snijders, 2007; McCulloh, Lospinoso and Carley, 2007; Lospinoso, 2008). One recent model involved a new node randomly attaching to the network and then "burning" through the network making new connections with other nodes like a wildfire spreads (Leskovec, Kleinberg, and Faloutsos, 2005). Unfortunately, these models do not leave us with a tractable analytic distribution of degree.

This chapter will highlight three novel points relevant to network science in general and this thesis in particular. First, different statistical distributions are fit to the same simulated data, showing that it is extremely difficult to assert with any certainty the

actual distribution of network properties. In other words, there are several candidate distributions which may all fit the data equally well. Second, the context that defines a relationship may significantly affect the structure of the network. For example, a social network where a relationship is defined as person *i* has seen person *j* before will be much more dense than a network where a relationship is defined as person *i* spends at least 1 hour in conversation with person *j* each day. Third, a power series distribution is proposed for the degree measure in a network. The binomial distribution common in random networks is shown to be a special case of the power series distribution. The power series distribution is also a power law distribution, which can fit the distribution of degree in other network structures. The parameters of the power series distribution can be viewed as a constraint function on the utility and costs associated with establishing a link. In the previous example, a node incurs greater cost in terms of time to spend one hour per day with another node than the cost associated with simply seeing another node one time. Together these three points provide direction for future research and identify an area of caution for detecting change which will be articulated later in this thesis.

2.1 Random Networks

The simplest model of a network is the *random network* (Erdos and Renyi, 1959, 1960, 1961). This model contains a fixed number of nodes, n. Between each ordered pair of nodes, a link is drawn with some probability, p. Under this model, all nodes have an equal probability of being connected to every other node and therefore they behave similarly to one another. Individual nodes can occupy very different positions in the network with the same random probability. Some nodes will be on the periphery of the network, while others will occupy a more central position. The degree to which this occurs depends only upon the random realization of an instance of the network.

If we fit statistical distributions to the nodal measures in a random network, we can create a statistical model to evaluate the position nodes occupy in the network. If a node occupies a position that is significantly anomalous to positions expected under random network assumptions, we might conclude that the network has an unusual node or that the topology of the network is not random. This insight is the first step towards a better statistical understanding and description of network behavior.

The degree, k_i, of node i is defined as the number of other nodes directly linked to i. The degree of a node will contribute to their influence in the overall group (Freeman, 1977). The degree will not determine influence alone, however. In Figure 4, it is clear that nodes 5 and 9 have the highest degree, but node 10 may have greater influence in the network due to its' more central position (Freeman, 1977; Wasserman and Faust, 1994).

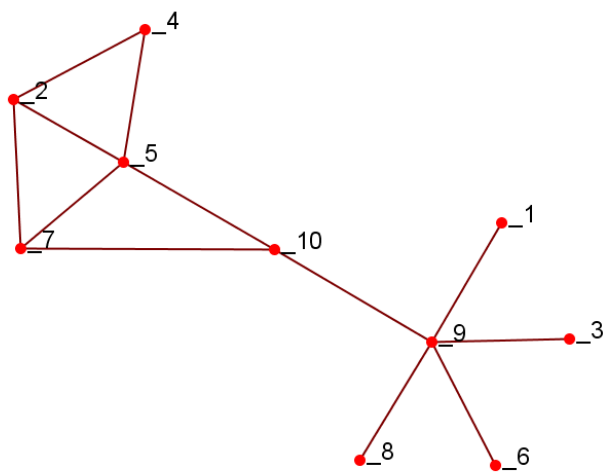

Figure 2. Example Network.

The distribution of degree in a random network has been well established in the literature (Erdos and Renyi, 1960; Donninger, 1986; Bollobas, 1998; Caldarelli, 2007). Suppose a network with n nodes follows the Erdos and Renyi model with probability p. Then, a node of degree k occurs if and only if there are exactly k link successes of the $n - 1$ possible links. The degree k follows a binomial distribution, satisfying the four properties of the binomial distribution: 1) There are a fixed number of nodes, n, in the network; 2) There is a fixed probability, p, of any two ordered pairs of nodes being linked; 3) The trials are identical, which occurs when nodes behave similarly; 4) The trials are independent, which means that a node's choice of connections is not influenced by other nodes in the network. The binomial degree distribution is given by,

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} .$$

The average degree is therefore given by $p(n - 1)$ and the variance is given by $(n - 1)p(1 - p)$. The maximum likelihood estimate (MLE) of the average degree is given by

$$\bar{k} = \frac{1}{(n-1)} \sum_{i=1}^{n-1} k_i = p(n-1) .$$

Therefore, the MLE of $p = \bar{k} / n$. If we assume that the node degrees are independent[1], then by the Central Limit Theorem, for large n, the average degree is an approximately normally distributed random variable. Since the probability of link occurrence, p, is a linear combination of a normal random variable, it too is a normal random variable. We can now proceed to derive a confidence interval about the estimate of p. This is done

[1] This assumption is not valid for all networks. The impact of ergodicity in networks is still an open research area.

more easily by relating the distribution of the number of links in the network to the parameter p.

We can derive a similar distribution for the number of links, L, in a random network. There are $n(n-1)$ possible links in a social network if we exclude reflexive links. If the probability of a link occurring is a constant probability, p, then the distribution of links in a random network also follows a binomial distribution given by,

$$P(L) = \binom{n(n-1)}{L} p^L (1-p)^{n-1-L}.$$

If the number of links follows a binomial distribution, then the average number of links is given by $np(n-1)$ and the variance of the links is given by $np(n-1)(1-p)$. The MLE of p is therefore $\bar{p} = L / n(n-1)$. Recall that we concluded that p is a normally distributed random variable when n is large and node degree is independent. If we take the difference of a normally distributed random variable and its' MLE and divide the result by the standard deviation of that random variable, we have a standard normal random variable. Therefore,

$$z_L = \frac{P - \hat{P}}{\sqrt{\hat{P}(1-\hat{P}) / n(n-1)}} \sim \text{Normal } (0,1)$$

This algebraically reduces to $P = \hat{P} \pm z_{L\alpha/2}\sqrt{\hat{P}(1-\hat{P}) / n(n-1)}$, which defines a confidence interval on $\bar{p}$ at the $1-\alpha$ confidence level, where $\hat{P} = L / n(n-1)$.

Any value of $L / n(n-1)$ or of $\bar{k} / n$ that exceeds the constructed confidence interval is therefore statistically anomalous at the $1-\alpha$ confidence level for a random network. If we conclude that a network does appear random, we can conduct a similar procedure for the degree to determine statistically anomalous nodes in the network.

$$z_k = \frac{k - \bar{k}}{\sqrt{np(1-p)}} = \frac{k - p(n-1)}{\sqrt{np(1-p)}} \sim \text{Normal } (0,1).$$

This reduces to $k = p(n-1) \pm z_{k_{\alpha/2}}\sqrt{np(1-p)}$. Any value of k that exceeds this value is statistically anomalous at the $1-\alpha$ confidence level for a random network.

2.2 Scale-Free Networks

Scale-free networks were introduced by Barabasi and Albert (1999). The term scale-free refers to their observation that "the distribution of their local connections is

free of scale, following a power law." The density function of the power law distribution used by Barabasi and Albert is given by,

$$f(k) = ck^{-\gamma},$$

where c is a positive finite constant and $\gamma > 0$. The parameter γ can be referred to as the shape parameter and c can be referred to as the scale parameter. The particular range of values for γ that makes a network "scale-free" has not been defined (Bollobas and Riordan, 2003; Albert, 2008; Alderson, 2008; Barabasi, 2008). This is an important distinction, because a value of γ that approaches 0 approximates a uniform distribution, which has very different properties than the networks we typically think of as scale-free. At the U.S. Military Academy's 3rd Network Science Workshop, Barabasi stated, "the interesting networks that we look at have a [shape parameter] between 2 and 3." For these networks the mean and higher order moments of the degree distribution are undefined. When the moments of a distribution are undefined, it becomes difficult to characterize the modeled process in a compact form. This is perhaps one of the intriguing challenges presented by modeling network structure.

Scale-free networks present a much more fundamental problem in terms of degree distribution. So far the networks most commonly modeled with a scale-free network are dichotomous, large networks with hundreds to millions of nodes. The degree of a node in a dichotomous network is always a discrete, countable, integer. Therefore, the power law density function applied in the network science community is merely an approximation of what is really a discrete probability mass function. While this approximation might be totally acceptable for networks with many nodes, keeping the distribution discrete allows us to explore the relationship between the random and scale-free networks.

A discrete version of the power law distribution is a power series distribution and is given by,

$$p(k) = \frac{a_k \gamma^k}{\sum a_k \gamma^k},$$

where the coefficient function $a_k > 0$, and γ is the shape parameter of the distribution. It is also possible to use a Zipf distribution (Newman, 2005), which is more analogous to the power-law distribution used by Barabasi and Albert. Both distributions follow a power law and have the ability to model data with a similar shape as I will show later in the chapter. I choose to use the power series distribution to show the similarity between random and scale-free networks. Both the Zipf and power series distribution are more appropriate distributions for modeling the degree of a node in a dichotomous network with a fixed number of nodes. Using the power series distribution, I show an interesting variate relationship that relates scale-free and random networks.

2.3 Variate Relationship Between Binomial and Power Series

The following derivation shows an interesting relationship between random and scale-free networks. As shown earlier, the degree distribution in a random network follows a binomial distribution. The degree distribution in a scale-free network follows a power series distribution. Binomially distributed data may appear as a discrete power series distribution when p is small (Evans Hastings and Peacock, 2000). As I stated earlier, the power series distribution is a more appropriate distribution for node degree in a dichotomous network, since the degree of a node is a countable integer bounded between 0 and $n - 1$.

The variate relationship between the binomial and the power series occurs where $k \sim$ Binomial(n, p), then $k \sim$ PowerSeries(γ), where the Power Series distribution is defined as,

$$\Pr(k) = \frac{a_k \gamma^k}{\sum a_k \gamma^k}$$

when $\gamma = p / (1 - p)$ and $\sum a_k \gamma^k = (1+\gamma)^n = (1-p)^{-n}$. Substituting these values into the expression for the power series probability mass function above, the mass function can be expressed as,

$$\Pr(k) = \frac{a_k (p/(1-p))^k}{(1-p)^{-n}} = a_k p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}.$$

The resulting mass function is the binomial probability mass function. It is important to note that the power series distribution can also be used to represent non-binomial distributed data by modifying the coefficient function, a_k. This modification allows the power series distribution to model the degree distributions commonly associated with scale-free networks. However, this variate relationship presents a single distribution that can be used to model both the degree distribution found in random and scale-free networks.

A limitation of the binomial is that the expected degree is always $n*p$, which can become unrealistic in very large networks. It can also be shown that for a critical value of p the network will undergo a phase transition between a network with many unconnected components to a single, giant component (Bollobas, 1998). These two behaviors typically distinguish random networks from scale-free networks.

2.4 Constrained Node Network

The random network is not always an unreasonable model for the distribution of degree in social networks (McCulloh et al, 2008; Ring, McCulloh and Henderson 2008). This is especially true of small networks under 100 nodes. I have conducted an extensive

search of the literature and modeled the degree distribution of many social networks (Newcomb, 1961; Sampson, 1969; Freeman and Freeman, 1980; Krackhardt, 1987; Johnson, Boster and Palinkas, 2003; McCulloh et al, 2007). When the relationship between nodes requires a meaningful investment of time, the social networks are not scale free. A person's website on Facebook is unconstrained in the sense that it can maintain an unlimited number of links pointing to it (in-degree). It is slightly constrained in the number of links that it points to (out-degree) in that it takes some amount of time and effort to find friends and establish a link to their web page. When we stipulate that two individuals must spend an hour together and engage in meaningful conversation each week for a relationship (link) to exist, we have severely constrained the degree of the node. The scarce resource of time prevents any node from maintaining a large number of connections as they might in a scale free network.

The size of the network is also an important consideration in degree distribution. A node's choice of which other nodes to connect with can be limited by the size of the network. For a group of 20 individuals, for example, it is quite possible that all people have a connection. When the size of the group contains thousands of individuals, it becomes virtually impossible for all nodes to be connected. Simply consider the amount of time required to hold a meaningful conversation between a given node and all others in the network. This further suggests that there is some kind of limitation on the number of connections a node can have. In addition to the cognitive limitations of a person, nodes can be constrained by proximity, the utility of the connection, and in some cases the cost of the connection.

The constraints on link formation between nodes in a network will be context specific. The number of social ties that a person can maintain might be limited by their time, cognitive capacity, proximity, and other factors. A web-page, on the other hand, is not limited in the number of other web pages that can connect to it. There are limitations, or at least a cost in terms of building the web site, in the number of links to other pages, and still other factors. Relationships in a social network can also be very general, such as "been to the same country", which would connect many individuals that do not know each other at all.

The probability of link formation could therefore be modeled as a function of the context specific constraints. The constraints could prevent the occurrence of hub nodes, characteristic of scale-free networks in one context, but not in another. While the functional forms could be similar between different contexts, and therefore their network structure will be similar; the underlying functions governing constraints in the network may be quite different. Therefore, I propose modeling the constraints in the network as some form of the coefficient function a_k. An understanding of how network constraints can affect topological properties within the network is dependent upon how relationships are defined in network data. This is critical to understanding the behavior of the network.

2.5 Virtual Experiment

A virtual experiment is conducted to show how power series distributions with differing constraint functions can be used to model scale-free networks generated under the Barabasi-Albert model. I remind the reader that real-world data does not follow this model for many applications, particularly when the relationships are constrained by time or other scarce resources. However, scale-free networks have been successfully used to model the internet, electrical networks, protein networks, and other real-world data (Goh, Kahng and Kim, 2001; Pastor-Satorras and Vespignani, 2004; Leskovec, Kleinberg and Faloutsos, 2005; Caldarelli, 2007). Since random networks are already an established model in the social network community (Wasserman and Faust, 1994), I compare the power series distribution to the scale-free network in an attempt to unify the competing models.

In order to demonstrate the flexibility of the power series distribution and demonstrate differing constraint functions, 3000 scale-free networks were simulated consisting of 1000 nodes each. Networks were generated with degree distributions having $\gamma = 2$, 2.5, and 3. These values of γ were chosen because they were said to be the "interesting" networks by Barabasi (2008). Since he coined the term "scale-free" network, these seemed to be the best point of comparison found in the literature. I generated 1000 networks for each value of γ, therefore there were 3000 total scale-free networks.

Four statistical distributions are fit to the degree of the networks using the method of least squares. This method is chosen, because it minimizes the error in the fit of the distribution, which can be used to compare the quality of the fit among the four different distributions. The four distributions include the Barabasi and Albert power law distribution and three power series distributions, each with a different constraint function. The first constraint function is the *binomial constraint* function given by, $a_k = \binom{n-1}{k} = \binom{999}{k}$. The second power series distribution uses a *constant constraint* function given by, $a_k = 1$. The third power series distribution uses a more flexible *inverse constraint* function given by, $a_k = \left(\frac{1}{k}\right)^{\lambda}$.

The coefficient of determination is calculated for each distribution's fit to each of the 3000 networks. The coefficient of determination is calculated using the formula,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{p}_i - p_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p}_i)^2},$$

where p_i is proportion of the empirical data that has a degree less than the degree of node i, $\hat{p}_i$ is the probability of observing a random value less than the degree of node i under proposed distribution, and $\bar{p}_i$ is the average of the proportion of the empirical data which is equal to 0.5. As the error in the fit of the distribution increases with respect to the variance, the value of R^2 decreases. Using the method least squares will minimize the error in the fit of the distribution. Since we are comparing which distributions are appropriate for the data, this method maximizes the value of R^2 and gives each distribution a best case fit for comparison. This is particularly important with heavy tail data such as the power-law, since observed values in the tail of the distribution can

significantly bias the estimation of distribution parameters. I acknowledge that the parameters of those distributions are not minimally variant unbiased estimators for the distributions. In practice, it is desirable to use a Maximum Likelihood Estimate of the distribution parameters. For this application, the method of least squares is clearly more appropriate.

The coefficients of determination were compared between a Barabasi - Albert power law distribution fit to the data, and three power series distributions with different constraint functions fit to the same data. This was repeated for all 3000 networks. A two-sample t-test is used to evaluate the null hypothesis that the coefficients of determination of the best fit power series distribution are the same as the Barabasi - Albert power law distribution. An important distinction is made. In this virtual experiment, the underlying stochastic process that generates the data is the Barabasi - Albert power law distribution (Albert and Barabasi, 1999). The coefficient of determination for this distribution is therefore a measure of the error in the simulated data. When there is insufficient evidence to reject the null hypothesis, this only suggests that the power series distribution is a possible alternative distribution. In real-world data, we cannot know the underlying distribution. Therefore, a lack of evidence to reject the null hypothesis in this virtual experiment has important implications for empirical studies. It means that there is no evidence to suggest that the Barabasi - Albert power law distribution is a better fit to the empirical data than the best fit power series distribution.

2.6 Results

The 1000 scale-free networks for each parameter of γ were successfully created for a total of 3000 independently seeded networks. The node degree for all 1000 nodes in each network was recorded. Figure 3 displays the empirical distribution function for a representative instance of the degree distribution for the networks where $\gamma = 2$, 2.5, and 3 respectively. This figure illustrates the importance of the parameter γ in defining a scale-free network. As the value of γ approaches 1, a uniform distribution can be fit to the data. As the value of γ increases, the curvature in the data increases.

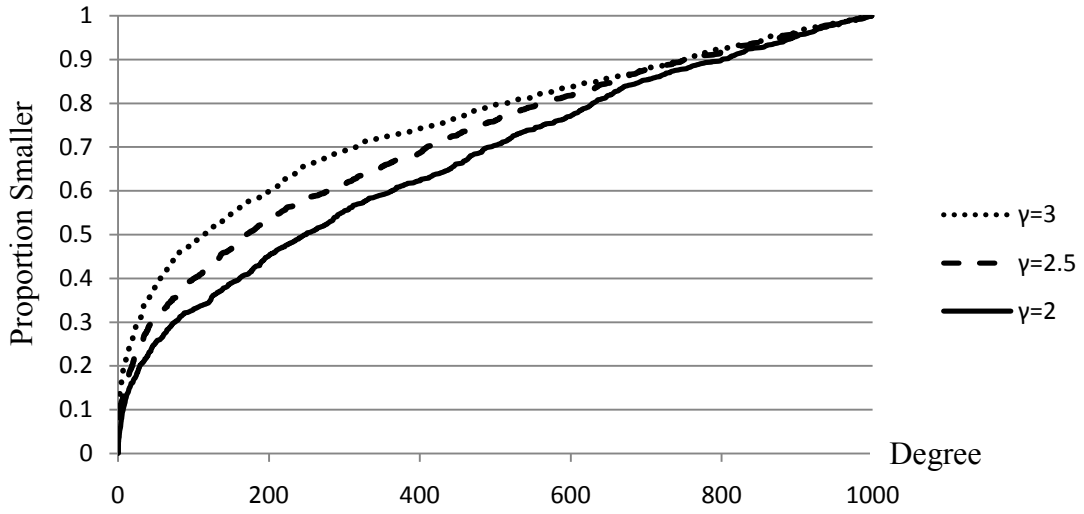

Figure 3. Empirical distribution functions for the degree of a scale-free network

The three power series distributions and the Barabasi-Albert power law distribution were fit to each of the 3000 generated networks. Figure 4 shows a representative instance of the empirical distribution, each of the three power series distributions using the different constraint functions, and the Barabasi-Albert power law distribution fit to the most extreme degree distribution where $\gamma = 3$. There is a noticeable difference between the fit of the power series distribution using the binomial constraint function and the empirical distribution function. This has been extensively noted in much of the network science literature. There is not a noticeable difference between the power series distribution using the inverse constraint function and the empirical distribution function. This suggests that a power series distribution may provide an equally good fit to network data as the Barabasi-Albert power law distribution provides.

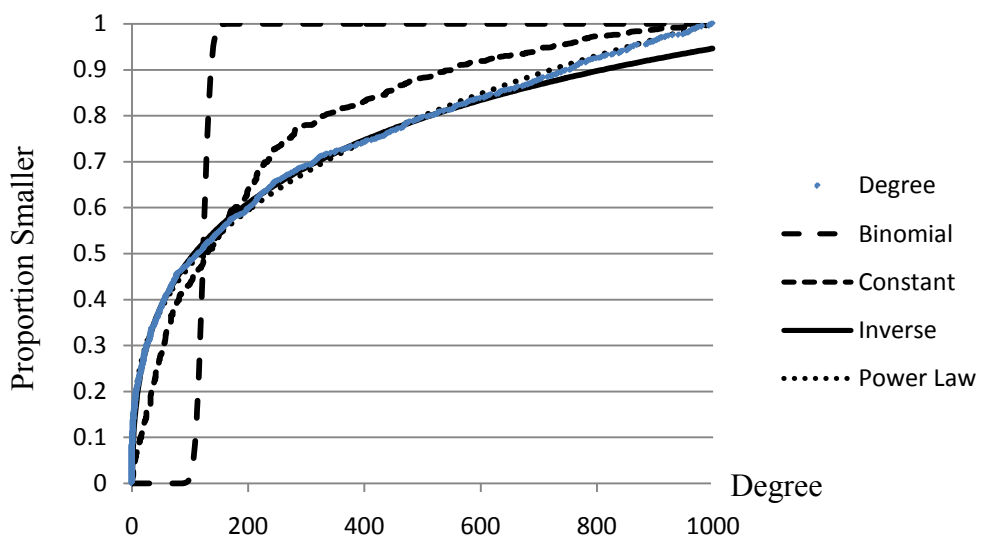

Figure 4. Various distributions fit to scale-free empirical degree distribution.

The coefficients of determination provide a quantitative measure of the quality of fit of the distribution. Table 2 shows the average coefficients of determination across the 1000 simulated networks of each parameter value γ, for the four distributions fit to each empirical distribution function. A value of 1.0 indicates that the distribution fits the data with no error. A value of 0.0 indicates that the distribution has no explanatory power in describing the data.

Table 2. Coefficients of Determination for Four Distributions.

	Barabasi-Albert	Binomial	Constant	Inverse
$\gamma = 2.0$	0.9984	0.0929	0.9444	0.9970
$\gamma = 2.5$	0.9989	0.0927	0.9111	0.9971
$\gamma = 3.0$	0.9966	0.0926	0.8531	0.9937

The power series distribution with the binomial constraint function does not fit the empirical data well. It is possible to improve the fit of this distribution if I allowed the size parameter, n, to increase. Since the size of the network was set at 1000 nodes, I decided it would be a more appropriate comparison to fix this parameter at $n = 1000$ to be consistent with the other distributions. The power series distributions with both the constant and inverse constraint functions provide a reasonably good fit to the data. I again remind the reader that the values in Table 2 compare the power series distributions to an empirical data set that definitely has a Barabasi-Albert power law underlying stochastic process. It has not been established that the Barabasi-Albert power law distribution models real-world empirical data as well as the power series with constant constraint function. In fact there have been several papers that have questioned the appropriateness of the Barabasi-Albert power law distribution applied to empirical findings (Doyle et al, 2005; Wasserman, Scott and Carrington, 2007; Alderson, 2008).

The power series distribution with the inverse constraint function is compared to the Barabasi-Albert power law distribution using a two sample t-test. The t-test is repeated for each set of 1000 networks corresponding to $\gamma = 2$, 2.5, and 3. Table 3 shows the results of the t-test.

Table 3. Statistical Comparison of Power-Law and Power-Series Distributions.

	Barabasi-Albert	Inverse Constraint	Test Statistic, T	p-Value
$\gamma = 2.0$	0.9984	0.9970	0.933	0.3510
$\gamma = 2.5$	0.9989	0.9971	1.294	0.1960
$\gamma = 3.0$	0.9966	0.9937	1.611	0.1075

There is no p-value < 0.10 that would indicate a statistically significant difference at the 90% confidence level. Therefore, we do not have enough evidence to reject the null hypothesis that the distributions are different. The reality of this virtual experiment is that they are different. This can be seen in the marginal trend, where each of the coefficients of determination for the power series with inverse constraint function are less

than the true underlying distribution. However, the fact that there was not a statistically significant difference with 1000 simulation replications, suggests that the two distributions are suitable to model the same data. It can also be seen in Table 3 that the p-value has a decreasing trend as γ increases. The power series distribution may not be appropriate for modeling data generated with a Barabasi-Albert power law distribution with $\gamma > 3$.

2.7 Discussion

I believe that the power series family of distributions may be a more appropriate statistical distribution for modeling the degree distribution of nodes in network data. Selecting a statistical distribution based purely on empirical fit can be misleading. Certain statistical distributions have properties that can have implications for the underlying mechanics of the system being modeled. For example, the exponential distribution has a memoryless property. This implies that future observations are independent of past observations. While this is appropriate for modeling the inter-arrival time of customers at a bank, it can be very misleading when modeling the arrival time of a bus, which comes at regular intervals. While the application of waiting for arrival is the same, the underlying mechanics are very different. In one situation the time between arrivals is independent of the past. In the other situation, the time past since the last bus arrived holds a great deal of information about when the next bus will arrive. A similar analogy may be true for the degree distribution of a network.

It is therefore necessary to consider the fit of various candidate power-series distributions when investigating the distribution of degree in network data. If a power-series distribution can be fit reasonably well to the data, the nature of the coefficient function may provide some insight into constraints on the nodes within the network. Further investigation into the coefficient functions of various network data may provide greater insight into the behavior of networks and possibly lead to predictive network models. At the very least, a single family of distributions for degree may lead to understanding appropriate statistical distributions for other network measures that will allow greater hypothesis testing in network data.

The power series distribution's coefficient function has been suggested to represent constraints on node degree. Further investigation is warranted. I have only suggested that this is a possible explanation for network degree distributions and shown some interesting variate relationships. Perhaps multi-agent simulations could constrain individual node degree and provide insight into the constraint function. Constraint functions could be modeled for various real-world network data and inference could be drawn about constraints based on the known or assumed behavior of the application.
We must be careful about the conclusions we draw from the degree distribution of network data. Just because a particular candidate distribution fits empirical data well, does not mean that the underlying distribution generating the data is the same. Scale-free networks that have a power law distribution were not necessarily evolved through preferential attachment (Albert and Barabasi, 1999) as some may assume. Preferential

attachment will produce scale-free networks, but the observation that the network is scale-free does not allow us to infer how it was created.

The power series distribution offers an alternative model to the Barabasi-Albert power law distribution. I believe that this is an appropriate statistical model due to the explanatory nature of the constraint function in the distribution. There is currently no intuitive explanation for the occurrence of power-law distributions in networks. In some sense the statistical support of the binomial distribution is more accurately defined for degree distribution than a continuous power-law distribution, due to the countable, integer values that degree can assume. Hopefully, future researchers will explore the topology of network data, looking at a wider range of candidate distributions and the implications for science that they may uncover.