Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

MEASURING THE EFFECTIVENESS OF THE S-METRIC TO PRODUCE BETTER NETWORK MODELS

Isabel Beichl
Brian Cloteaux

Mathematical and Computational Sciences Division
National Institute of Standards and Technology
100 Bureau Drive, Stop 8910
Gaithersburg, MD 20899-8910, U.S.A

ABSTRACT

Recent research has shown that while many complex networks follow a power-law distribution for their node degrees, it is not sufficient to model these networks based only on their degree distribution. In order to better distinguish between these networks, the metric s was introduced to measure how interconnected the hub nodes are in a network.

We examine the effectiveness of creating network models based on this metric. Through a series of computational experiments, we compare how well a set of common structural network metrics are preserved between instances of the autonomous system Internet topology and a series of random models with identical degree sequences and similar s values. We demonstrate that creating models based on the s metric can produce moderate improvement in structural characteristics over strictly using degree distribution. Our results also indicate that some interesting relationships exist between the s metric and the various structural metrics.

1 INTRODUCTION

A type of data of increasing importance in various areas of research is one that is based not on floating point values but rather on connections between objects that can be modeled as massive graphs or networks. Creating realistic models of these systems is necessary to understanding the interactions involved.

Unfortunately many real world systems cannot be approximated using simple random graphs. The reason is that the degree distribution of an Erdős-Rényi random graph follows a Poisson distribution. The number of edges connected to a node is called the *degree* of the node and the set of all the degrees in a graph is called its *degree distribution*. For many complex systems however, it has been shown (Barabási and Albert 1999) that the degree distribution of the resulting networks follows a power law distribution. In other words, the probability that a node has k adjacent edges is $P(k) \sim k^{-\alpha}$ for some $\alpha > 1$. This distribution in a graph produces a few nodes with very high degree (often called *hub nodes*) and a large number of low degree nodes. An example of this type of distribution is seen in Figure 1.

The importance of networks having power law distributions lies in the number of application areas in which they are found. These networks have been shown to arise naturally in systems of both biological (Watts and Strogatz 1998, Jeong et al. 2000, Williams and Martinez 2000) and social (Amaral et al. 2000) interactions. They also appear in many engineered systems such as the power grid (Watts and Strogatz 1998), the Internet (Faloutsos, Faloutsos, and Faloutsos 1999), and software components (Potanin et al. 2005). Thus, realistic models of these type of interactions need to reflect their power law distribution.

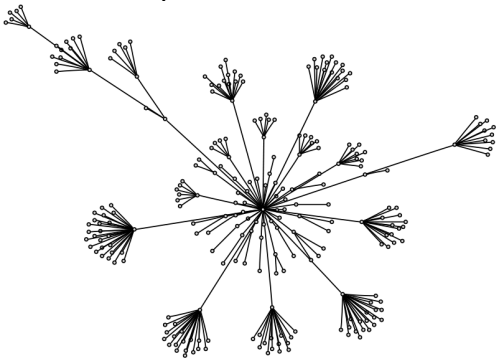


Figure 1: This figure shows a spanning tree of a subset of the autonomous system domain of the Internet. This is an example of a power law like degree distribution with a few high degree nodes and many nodes of degree 1 or 2.

However, even generating random graphs using power law distributions is still not sufficient to model real networks. Informally, the reason for this discrepancy is that for any given degree sequence there can be a large number of non-isomorphic graphs that share that particular degree sequence. Thus we can expect to see a large variability in the characteristics of the graphs that share a degree sequence.

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

In order to address the problem of distinguishing graphs with the same degree sequence, Newman (2002) introduced the concept of the *assortative mixing* of a graph. Assortative mixing is a measure of the preference for high degree nodes in network to attach to other high degrees nodes. The importance of this measure is based on two observations: one, that in many real networks there is a high affinity for high degree nodes to connect to each other and two, differences in the amount of assortative mixing can substantially affect the characteristics of these networks.

This idea of measuring how the nodes of high degree are connected to each other was further developed by Li et al. (2005). In their paper, they introduce the s metric as an alternative for measuring the connectedness of high degree nodes. The advantages of their metric is that it is simple to compute and it has been shown to be able to distinguish between many graphs with identical degree sequences. We will give a precise definition of the s metric later in the paper.

The purpose of this paper is to examine how well random networks generated with a given s value are able to model the structural characteristics of real networks. We performed a set of experiments by taking an actual network, the topology from the autonomous system (AS) domain of the Internet, and measuring if random graphs with identical degree sequences and similar s values maintain similar structural characteristics. Our conclusion is that the models produced using the s metric are superior to those using simple uniform sampling of graphs with the same degree sequence. While in most cases, the improvement was modest, the s metric seems to be an important first step towards understanding how the structure of these networks affects their capabilities. We first define the s metric and explain the methodology of our experiments. We then show how well a series of structural metrics were preserved using these modeling constraints.

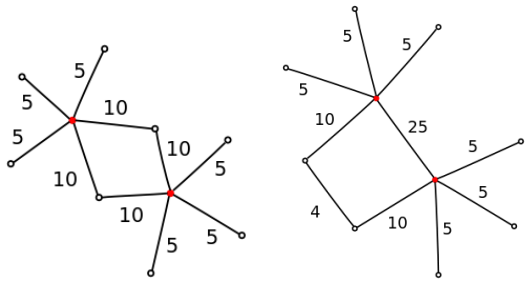
2 DEFINITIONS AND EXPERIMENTAL SETUP

2.1 The s metric

The s metric was proposed by Li et al. (2005) as a measure of interconnectedness between the hub nodes of a network. They showed that this metric is able to distinguish between many graphs with identical power law distributions. Before giving a definition of the s metric, we must first define some notation. A graph $G = (N,E)$ has a node set N and an edge set E. To show that an edge set E (or a node set N) belongs to a graph G, we write $E(G)$ (or $N(G)$). The *degree sequence* of G is defined as $\omega = \{\omega_1, \omega_2, ..., \omega_{|N|}\}$ where the degree of $n_i \in N(G)$ is ω_i. To show that a degree sequence ω belongs to a graph G, we write $\omega(G)$. The set of all simple connected graphs with the degree sequence ω is represented as $\mathscr{G}(\omega)$. The definition of the s metric for a graph G is

$$s(G) = \sum_{(i,j)\in E(G)} \omega_i \cdot \omega_j$$

In Figure 2, we see an example of how the structure of the hub nodes affects the s value of a graph.



(a) s value is 70 (b) s value is 79

Figure 2: This example shows two graphs with identical degree sequences but with differing s values. This illustrates that the more connected the hub nodes (in this case, the filled red nodes in the graphs) of a network are to each other, the greater the s value will be.

The S metric was also introduced by Li et al. (2005) as a normalization factor to the s metric. This is defined by

$$S(G) = \frac{s(G)}{s_{max}(\omega(G))}$$

where $s_{max}(\omega) = \max\{s(G)|G \in \mathscr{G}(\omega)\}$. This normalization factor allows us to compare networks with differing degree sequences. Otherwise, the s metric can only be profitably used to compare networks with identical degree sequences.

For the purposes of this paper we will ignore the S metric and concentrate only on the s metric. This simplification is for two reasons: first all the graphs we are comparing have identical degree sequences and so no normalization is needed, and second the computation of the s_{max} value is nontrivial for larger degree sequences. The measured s values for the different AS instances is given in Table 4(a).

2.2 Methodology

Our study is a set of computational experiments involving the degree sequences of three different instances of the AS topology. These three instances are approximately spaced two years apart at the dates January 1, 2004, November 25, 2005, and November 15, 2007. The data for the AS topology we used came from the UCLA Internet topology

collection (UCLA 2008). The size of the instances is shown in Table 3(a).

Using each instance of the AS topology, we generated three separate sets of graphs each having identical degree sequences as the AS topology. The size of all of the sets generated is given in Table 3(b).

AS Topology	Num. Nodes	Num. Edges
01/01/2004	16573	39143
11/25/2005	24567	103655
11/15/2007	32821	169106

(a) Size of AS instances

	Num. Graphs
01/01/2004 uniform	200
01/01/2004 s spectrum	127
01/01/2004 s model	51
11/25/2005 uniform	819
11/25/2005 s spectrum	167
11/25/2005 s model	50
11/15/2007 uniform	180
11/15/2007 s spectrum	163
11/15/2007 s model	34

(b) Number of graphs in sets

Figure 3: Description of the data sets

The first set (designated as the *uniform* set) contains random graphs sampled with an almost uniform probability distribution. We generated these graphs using the sequential importance sampling method from Blitzstein and Diaconis (2006). The program we used is a rewrite of code originally developed by Blitzstein (2007) for generating connected random graphs with a prescribed degree sequence. In order to generate larger networks, we needed to increase the speed of this program. For our version, we converted the original code from the language R to C++ and included various optimizations.

The second set (designated as the *s spectrum* set) contains random graphs selected to cover a spectrum of s values and the third set (designated as the *s model* set) has random graphs generated to approximate the s value of the AS instance. In order to generate a random model with an approximate s value, we used a random walk over the space of connected graphs with identical degree sequences. A detailed discussion on the algorithms used in the construction of these two sets in given in Beichl and Cloteaux (2008). The mean s values of generated graphs is shown in Table 4(b).

3 METRIC COMPARISON

For our investigation we examine how well four structural metrics are preserved by graphs with the same degree sequence and similar s values. The choice of three of the

AS Topology	s Value
01/01/2004	3.104e+08
11/25/2005	2.329e+09
11/15/2007	6.230e+09

(a) s value for AS instances

	Mean	Std Dev.
01/01/2004 uniform	5.856e+08	1.062e+06
01/01/2004 s model	3.108e+08	5.166e+05
11/25/2005 uniform	3.778e+09	6.609e+06
11/25/2005 s model	2.302e+09	2.433e+06
11/15/2007 uniform	9.676e+09	1.257e+07
11/15/2007 s model	6.233e+09	6.266e+06

(b) Mean s value for the generated sets

Figure 4: s values of the data sets

metrics comes from the list of common network metrics mentioned by Tangmunarunkit et al. (2002). The fourth metric, number of spanning trees, comes from our own investigations into network reliability. We give a short description of each of the metrics and the results of our comparisons.

3.1 Diameter

The diameter of a graph G is the maximum length of all the shortest paths between any two nodes in G. Diameter is a rough measure of the expected size of paths in a network, since the diameter must be at least as large as the mean path distance. In other words, the smaller the diameter of a network, the smaller we expect the length of the path between any two nodes in the network to be. The diameters for the AS instances and the set means are given in Figure 5.

AS Topology	Diameter
01/01/2004	10
11/25/2005	8
11/15/2007	9

(a) Diameter of the AS instances

	Mean	Std Dev.
01/01/2004 uniform	14.2	1.44
01/01/2004 s model	12.1	0.88
11/25/2005 uniform	10.4	0.66
11/25/2005 s model	9.8	0.76
11/15/2007 uniform	9.8	0.68
11/15/2007 s model	9.1	0.70

(b) Mean diameter for the generated sets

Figure 5: Diameters of the data sets

In Figures 6, 7, 8, we see a comparison of the diameters for the set of graphs over the spectrum of s values. Examining these graphs, it appears that mean diameter is correlated to the s value with the diameter spiking upward after we reach the random mean. While we are unsure of the precise mechanism for this dependence, we can get an intuitive feel for why this result seems reasonable by observing that as the s value approaches s_{max}, the hub nodes are forced to completely connect to each other. As this happens, it forces nodes whose degree are in the tail of the distribution to connect with one another. This tends to produce long 'strands' in the network that can greatly increase its diameter.

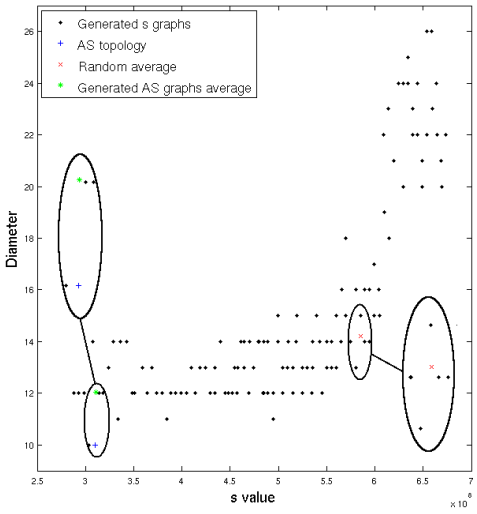


Figure 6: Diameters of the 01/01/2004 degree sequence

3.2 Biconnected components

A biconnected component $B \subseteq E(G)$ in a graph G is a maximal set of edges such that any two edges in the set are on some cycle in B. Since for any node in a biconnected component with size greater than one to be disconnected requires cutting at least two edges in the network, the number of biconnected components is a measure of how much edge redundancy a graph has. Generally, the smaller the number of biconnected components, the greater the number of paths between nodes in the network.

Computing the number of biconnected components can be accomplished in time linear to the number of edges in the graphs (Tarjan 1972). The number of biconnected components for the AS instances and the set means is given in Figure 9.

The results of our comparisons of the number of biconnected components over the s spectrum (Figures 10, 11, and 12) show that the s metric does a better job of modeling the biconnectivity of the AS topologies than simple random

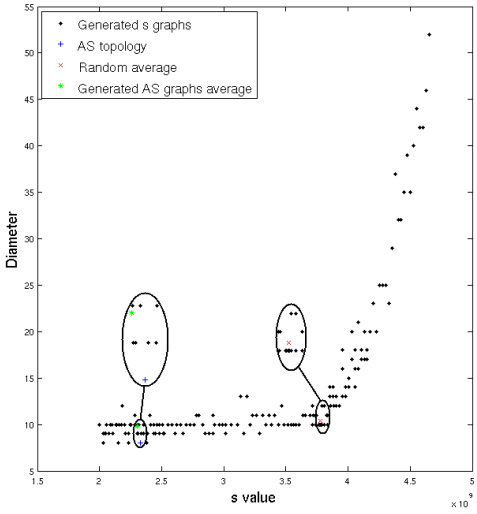


Figure 7: Diameters of the 11/25/2005 degree sequence

graphs. With regards the shape of the graph, we can note that diameter and the number of biconnected components are not uncorrelated. As the diameter increases, we start to see more *bridges* or edges that disconnect the graph if removed. Obviously, no biconnected component with more than one edge can contain a bridge, and thus we would expect the number of biconnected components to increase also.

We also notice that we see an 'elbow' in the graphs produced at the point of the random graph set's mean value. The fact that we consistently see this feature in the all graphs that we produced over all the different metrics seems remarkable. We do not have a complete understanding of this feature, and it remains an active area of our research.

3.3 Node cover

The minimum node cover of a graph G is a minimum set of nodes $NC(G)$ such that every edge in G is adjacent to at least one node in this set. The size of the minimum node cover is a measure of compactness of the network. We can think of this metric as measuring the smallest number of nodes we would need to monitor in a network to ensure the reliability of all the connections.

While the problem of finding a minimum node cover is NP-hard (Garey and Johnson 1990), there are efficient strategies for solving many specific instances. In particular, we used the kernelization techniques of Abu-Khzam et al. (2004) in order to simplify the problem. The idea of kernelization algorithms is to simplify the graph G to produce a smaller G' called the kernel of G. A constant k is also

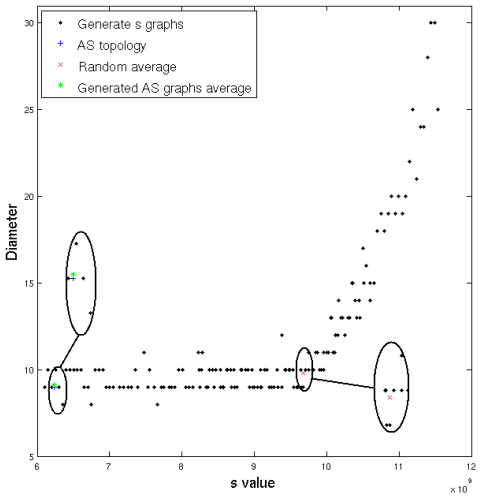


Figure 8: Diameters of the 11/15/2007 degree sequence

AS Topology	Biconnected Components
01/01/2004	5441
11/25/2005	3789
11/15/2007	4363

(a) Number of biconnected components for AS instances

	Mean	Std Dev.
01/01/2004 uniform	6627.0	34.2
01/01/2004 s model	6169.5	30.8
11/25/2005 uniform	4059.6	20.2
11/25/2005 s model	3971.2	16.2
11/15/2007 uniform	4552.9	16.5
11/15/2007 s model	4497.7	15.4

(b) Mean number of biconnected components for generated sets

Figure 9: Number of biconnected components for the data sets

given such that

$$|NC(G)| = |NC(G')| + k$$

In most of the graphs we examined, the resulting kernel graph was trivial to solve for its smallest node cover exactly. In many instances the resulting kernel was under 20 nodes and in some cases it was even empty. Thus, for almost all instances, we are able to compute exact values for this metric. The minimum node cover for the AS instances and the mean for the sets is given in Figure 13.

In Figures 14, 15, and 16, we see the behavior of the minimum node cover over the s spectrum. For some graphs with very large s values, the resulting kernel was too large for simple brute force computation. In those instances, as shown in the figures, we instead computed error bounds on the minimum node cover. The upper bound was created by minimizing the node cover using threshold acceptance. We derive a lower bound on the node cover using a well-known approximation algorithm to the problem that is guaranteed to produce a result that is never more than twice the minimum node cover (Vazirani 2001). We conjecture that the true minimum value for the graphs in those instances is actually much closer to the top of the error bars than the bottom.

The results of our comparisons of the minimum node cover over the s spectrum show that the s metric does improve over the uniform sampled graphs in modeling the minimum node cover of the AS topologies, but still does a poor job in the overall result. These real networks have much lower values than any network we were able to generate. Out of all the metrics we compared, this metric had by far the largest difference between the real networks and the generated models. Thus, the s metric does not seem to capture very well this particular notion of network compactness, but precisely why this happens is still unclear to us.

3.4 Spanning trees

A spanning tree in a graph is a set of edges such that each node is connected and there is no cycle in the edges. The number of spanning trees in a graph can be used as a measure of reliability. The reason is because, in general, as the number of spanning trees increases, the number of edges that need to be cut to disconnect the graph (or the *cut set*) also increases.

In order to efficiently count the number of spanning trees in a network, we have developed a Monte Carlo method based on sequential importance sampling. Our algorithm allows us to estimate the number of spanning trees of a graph along with all sub-forests with k edges for each k. While there does exist a polynomial time method for counting all spanning trees of a graph (for example, see chapter 1 of Jerrum (2003)), since it involves taking the determinant of the Laplacian matrix of a network it becomes impractical for the sizes of networks we are measuring. The estimated number of spanning trees for the AS instances and the mean for the sets is given in Figure 17.

When examining the graphs for the number of spanning trees in Figures 18, 19, and 20, we see a familiar pattern of the s models for the AS topology producing slightly better results than the simple random mean. We also see that the number of spanning trees rapidly decreases as we approach s_{max}. Again, we do not give a formal argument for this behavior, but we can try to explain it intuitively by relating it to the increase in biconnected components. As the number of biconnected components increase, the

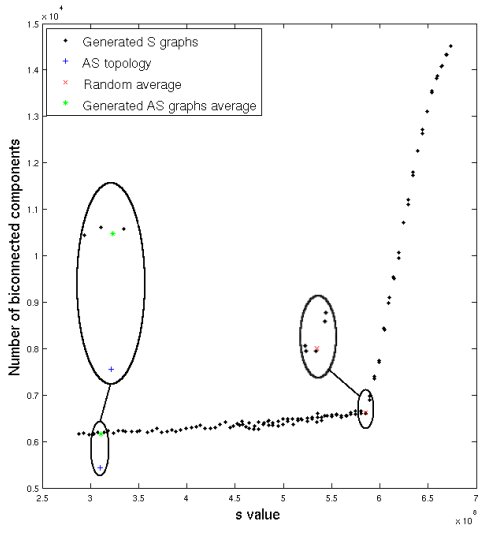


Figure 10: Number of biconnected components for the 01/01/2004 degree sequence

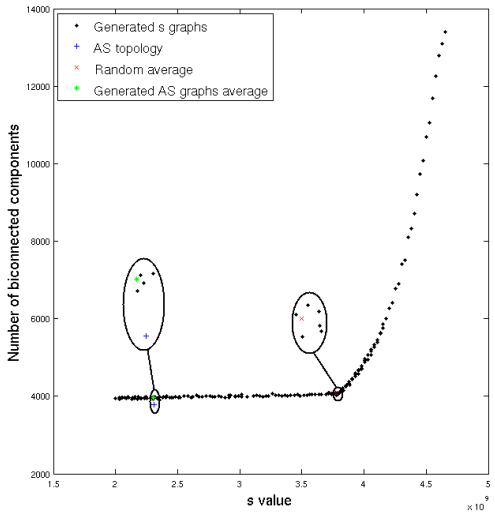


Figure 11: Number of biconnected components for the 11/25/2005 degree sequence

overall number of paths between the various nodes in the networks decrease. Since each spanning tree must contain a path between any two nodes, then the number of choices of edge combinations for constructing these spanning tree must decrease as well. We can also think about this relation as a result of the hub nodes all starting to become completely connected to each other. As this occurs, if we select $n-1$ edges at random from the graph, the probability that we have selected a spanning tree is lower than if all the nodes in the network are equally connected.

4 CONCLUSIONS

We started this investigation wanting to find better methods for creating better models of networks having power law distributions. Constraining the s metric of Li et al. (2005) has shown some promise in this regard. Using this metric, we were able to produce models with moderately better structural characteristics than uniformly selected random models. Where there are large differences between s metric generated graphs and real networks (such as the minimum node cover values), it has demonstrated that constraining the s value may be necessary for realistic modeling of these networks, but it is not sufficient. This has opened new lines of investigation into determining why these differences occur.

We also saw some unexpected outcomes in our experiments. In particular, we are continuing to try to understand the 'elbow' in the various graphs where the random mean exists. This has lead to us working on theoretical justifications for the shape of these graphs.

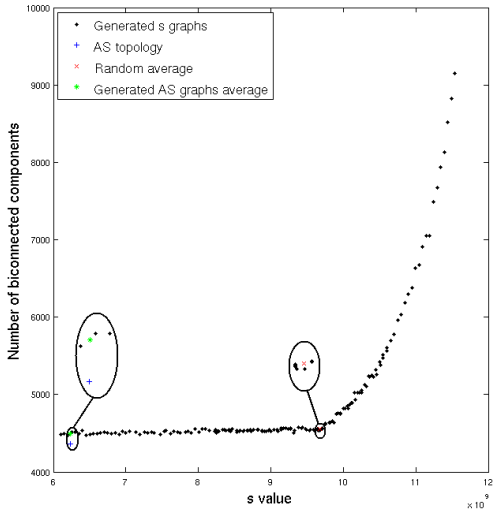


Figure 12: Number of biconnected components for the 11/15/2007 degree sequence

AS Topology	Min. Node Cover
01/01/2004	2465
11/25/2005	4898
11/15/2007	7058

(a) Minimum node cover for AS instances

	Mean	Std Dev.
01/01/2004 uniform	4932.0	29.7
01/01/2004 *s* model	4137.6	28.3
11/25/2005 uniform	8016.1	39.1
11/25/2005 *s* model	7018.4	38.0
11/15/2007 uniform	11045.0	41.9
11/15/2007 *s* model	9898.8	42.5

(b) Mean minimum node cover for generated sets

Figure 13: Minimum node covers of the data sets

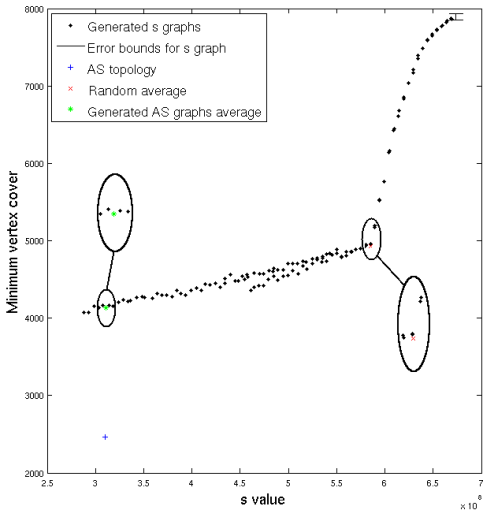


Figure 14: Minimum node covers for the 01/01/2004 degree sequence

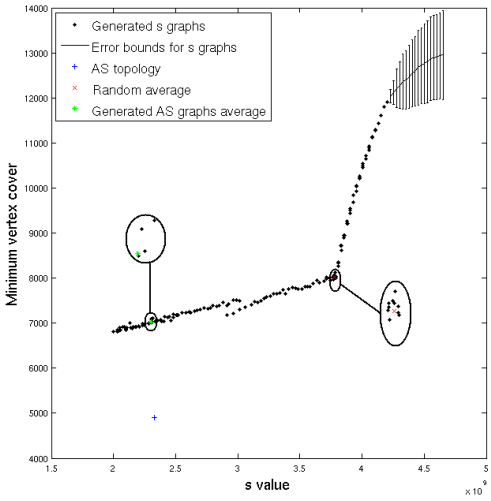


Figure 15: Minimum node covers for the 11/25/2005 degree sequence

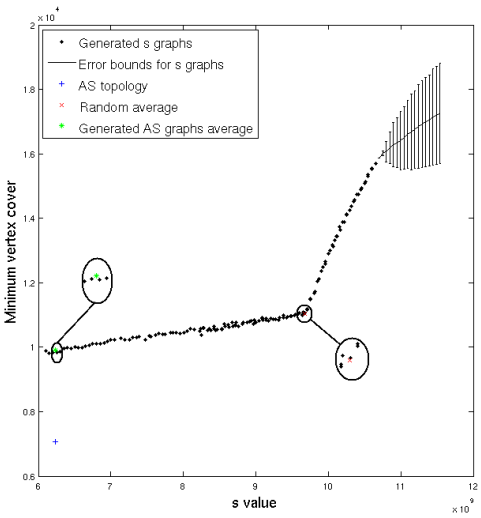


Figure 16: Minimum node covers for the 11/15/2007 degree sequence

AS Topology	Log Num Spanning Trees
01/01/2004	8511.52
11/25/2005	21192.2
11/15/2007	32205.7

(a) Natural log of the estimated number of spanning trees for the AS instances

	Log Aver.
01/01/2004 uniform	7768.5
01/01/2004 s model	7825.6
11/25/2005 uniform	20909.2
11/25/2005 s model	20943.5
11/15/2007 uniform	31292.4
11/15/2007 s model	31867.7

(b) Natural log of the means of the estimated number of spanning trees for the generated sets

Figure 17: Natural logarithm of the estimated number of spanning trees for the data sets

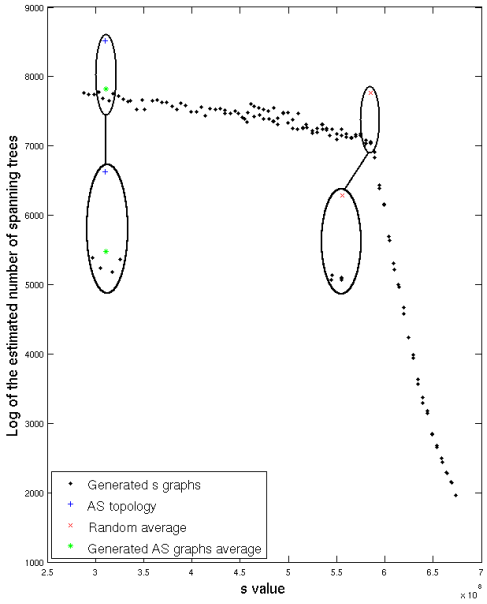


Figure 18: Log of estimated number of spanning trees for the 01/01/2004 degree sequence

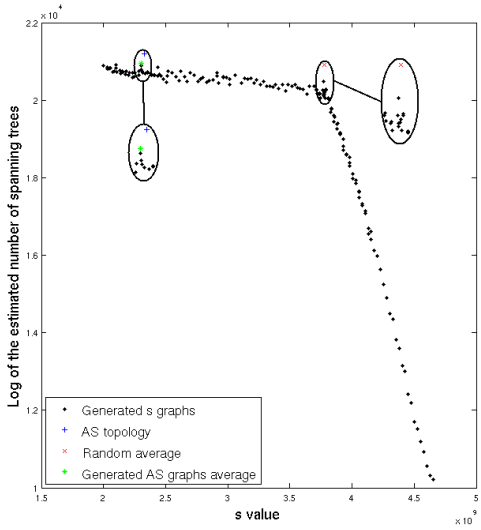


Figure 19: Log of estimated number of spanning trees for the 11/25/2005 degree sequence

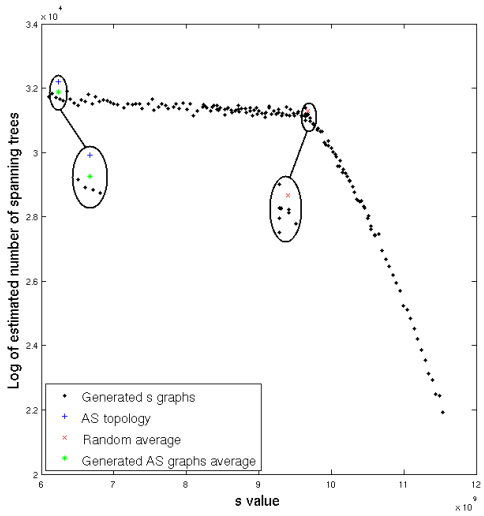


Figure 20: Log of estimated number of spanning trees for the 11/15/2007 degree sequence

Our conclusion on the *s* metric is that it is an important step to being able to quantify the structure of networks with power law distributions. Using this metric, there is a modest increase in model accuracy. An example application of this research is that we are currently looking to use these models to produce better simulations of the Border Gateway Protocol (BGP) routing systems of the Internet (Sriram et al. 2006). More importantly, understanding the relationship the *s* metric has to other structural metrics seems necessary to understanding how to create better methods for network characterization.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their helpful comments about this paper. We would also like to thank Ron Boisvert and Sandy Ressler for their suggestions on the presentation of the paper.

This work is a contribution of NIST, an agency of the US government, and is not subject to US copyright.

REFERENCES

Abu-Khzam, F. N., R. L. Collins, M. R. Fellows, M. A. Langston, W. H. Suters, and C. T. Symons. 2004. Kernelization algorithms for the vertex cover problem: Theory and experiments. In *Proc. 6th ACM-SIAM ALENEX*, 62–69: ACM-SIAM.

Amaral, L. A., A. Scala, M. Barthelemy, and H. E. Stanley. 2000, October. Classes of small-world networks. *Proc Natl Acad Sci USA* 97 (21): 11149–11152.

Barabási, A.-L., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Beichl, I., and B. Cloteaux. 2008. Generating network models using the *S*-metric. In *MSV '08: Proc. of the International Conference on Modeling, Simulation and Visualization Methods*, 159–164.

Blitzstein, J. 2007. Program for generating random graphs with prescribed degrees. http://www.people.fas.harvard.edu/~blitz/Site/Research_files/GraphAlgorithmR.txt. Accessed 10/12/2007.

Blitzstein, J., and P. Diaconis. 2006. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. Submitted to *Annals of Applied Probability*.

Faloutsos, M., P. Faloutsos, and C. Faloutsos. 1999. On power-law relationships of the internet topology. In *SIGCOMM '99: Proc. of the conference on applications, technologies, architectures, and protocols for computer communication*, 251–262. New York, NY, USA: ACM.

Garey, M. R., and D. S. Johnson. 1990. *Computers and intractability; a guide to the theory of NP-completeness*. New York, NY, USA: W. H. Freeman & Co.

Jeong, H., B. Tombor, R. Albert, Z. Oltvai, and A. L. Barabási. 2000, October. The large-scale organization of metabolic networks. *Nature* 407 (6804): 651–654.

Jerrum, M. 2003. *Counting, sampling and integrating: Algorithms and complexity*. Birkhauser.

Li, L., D. Alderson, J. C. Doyle, and W. Willinger. 2005. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics* 2 (4): 431–523.

Newman, M. E. J. 2002, Oct. Assortative mixing in networks. *Physical Review Letters* 89 (20): 208701.

Potanin, A., J. Noble, M. Frean, and R. Biddle. 2005. Scale-free geometry in OO programs. *Communications of the ACM* 48 (5): 99–103.

Sriram, K., D. Montgomery, O. Borchert, O. Kim, and D. R. Kuhn. 2006, October. Study of BGP peering session attacks and their impacts on routing performance. *IEEE Journal on Selected Areas in Communications* 24 (10): 1901–1915.

Tangmunarunkit, H., R. Govindan, S. Jamin, S. Shenker, and W. Willinger. 2002. Network topology generators: degree-based vs. structural. *SIGCOMM Comput. Commun. Rev.* 32 (4): 147–159.

Tarjan, R. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing* 1 (2): 146–160.

UCLA 2008. UCLA internet topology collection. http://irl.cs.ucla.edu/topology. Accessed 1/14/2008.

Vazirani, V. V. 2001. *Approximation algorithms*. Springer.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440–442.

Williams, R. J., and N. D. Martinez. 2000, March. Simple rules yield complex food webs. *Nature* 404 (6774): 180–183.

AUTHOR BIOGRAPHIES

ISABEL BEICHL is a mathematician in the Mathematical and Computational Sciences division at the National Institute of Standards and Technology. She holds a PhD degree in Mathematics from Cornell University. Her email address is <isabel.beichl@nist.gov>.

BRIAN CLOTEAUX is a computer scientist in the Mathematical and Computational Sciences division at the National Institute of Standards and Technology. He holds a PhD degree in Computer Science from New Mexico State University. His email address is <brian.cloteaux@nist.gov>.