CHAPTER 1

WEIGHTED SPECTRAL DISTRIBUTION: A METRIC FOR STRUCTURAL ANALYSIS OF NETWORKS

DAMIEN FAY[1], HAMED HADDADI[1,2], ANDREW W. MOORE[1], RICHARD MORTIER[3] AND STEVE UHLIG[4]

[1]University of Cambridge, UK
[2]University of London, UK
[3]University of Nottingham, UK
[4]Deutsche Telekom Laboratories/Technische Universität Berlin, Germany

1.1 SUMMARY

We consider the problem of structural comparison of graphs with a focus on a particular dynamic graph, the Internet's Autonomous System (AS) topology (§1.2). We develop the *weighted spectral distribution* (WSD), a metric based on the distribution of a particular decomposition of a graph's structure (§1.3) with a worked example (§1.4). We then turn to our particular application domain (§1.5), describing existing measures used to characterize Internet topologies, common topology generators, and several observed datasets used in our evaluation. We then compare the topology generators to the observed datasets using both existing measures and the WSD (§1.6), use the WSD to examine the impact of varying parameter selection for the different generators (§1.7), and optimize parameter values for the generators with respect to one of the observed datasets and examine the results using both WSD and traditional measures (§1.8). Finally we look briefly, from a particular vantage point, at the structural evolution of the Internet topology (§1.9), before concluding (§1.10).

(Machine Learning Approach for Network Analysis: Novel Graph Classes for Classification Techniques, First Edition). By (M. Dehmer and S. Basak)

1.2 INTRODUCTION

Graph comparison is a problem that occurs in many branches of computing, from vision to speech processing to systems. Many techniques exist for graph comparison, e.g., the edit distance [6] (the number of link and node additions or deletions required to turn one graph into another), or counting the number of common substructures in two graphs [21]. Unfortunately, these methods are too computationally expensive for large graphs such as the Internet topologies studied here. Moreover, they are inappropriate for dynamic graphs, resulting in varying edit distances or substructure counts. Common currently used "metrics" include the clustering coefficient, the assortativity coefficient, the node degree distribution and the k-core decomposition. However, these are not metrics in the mathematical sense, but rather are measures. This distinction is important as *a measure cannot be used to determine unique differences between graphs*: two graphs with the same measures may not in fact be the same. For example, two graphs may have the same clustering coefficient but hugely different structures.

In this chapter we present the *weighted spectral distribution* (WSD), a true metric in the mathematical sense, which compares graphs based on the distribution of a decomposition of their structure. Specifically, the WSD is based on the spectrum of the normalized Laplacian matrix and is thus strongly associated with the distribution of *random walk cycles* in a network. A random walk cycle occurs when we find we have returned to a node having walked N steps away from it. The probability of a random walk cycle originating at a node indicates the connectivity of that node: a low probability indicates high connectivity (there are many routes, few of which return) while a high probability indicates high clustering (many of the routes lead back to the original node).

The WSD is computationally inexpensive and so can be applied to very large graphs (more than 30,000 nodes and 200,000 edges). Also, it expresses the graph structure as a simple plotted curve that can be related to two specific properties of graphs: hierarchy and local connectivity. Given that the WSD is a metric in the mathematical sense several applications become possible: assessment of synthetically generated topologies based on real measurements, where the generated graphs should share some common structure with the original measurements rather than *exactly* matching them; parameter estimation for topology generators with respect to a target dataset; direct comparison among topology generators using these optimal parameters; and quantification of change in the underlying structure of an evolving topology.

1.3 WEIGHTED SPECTRAL DISTRIBUTION

We now derive our metric, the *weighted spectral distribution*, relating it to another common structural metric, the clustering coefficient, before showing how it characterizes networks with different mixing properties.

Denote an undirected graph as $G = (V, E)$ where V is the set of vertices (nodes) and E is the set of edges (links). The adjacency matrix of G, $A(G)$, has an entry of

one if two nodes, u and v, are connected and zero otherwise

$$A(G)(u,v) = \begin{cases} 1, & \text{if } u, v \text{ are connected} \\ 0, & \text{if } u, v \text{ are not connected} \end{cases} \tag{1.1}$$

Let d_v be the degree of node v and $D = diag(sum(A))$ be the diagonal matrix having the degrees along its diagonal. Denoting by I the identity matrix $(I)_{i,j} = 1$ if $i = j, 0$ otherwise, the Normalized Laplacian L associated with graph G is constructed from A by normalizing the entries of A by the node degrees of A as

$$L(G) = I - D^{-1/2}AD^{-1/2} \tag{1.2}$$

or equivalently

$$L(G)(u,v) = \begin{cases} 1, & \text{if } u = v \text{ and } d_v \neq 0 \\ -\dfrac{1}{\sqrt{d_u d_v}}, & \text{if } u \text{ and } v \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \tag{1.3}$$

As L is a real symmetric matrix there is an orthonormal basis of real eigenvectors $e_0, \ldots, e_{n-1}$ (i.e., $e_i e_j^T = 0, i \neq j$ and $e_i e_i^T = 1$) with associated eigenvalues $\lambda_0, \ldots, \lambda_{n-1}$. It is convenient to label these so that $\lambda_0 \leq \ldots \leq \lambda_{n-1}$. The set of pairs (eigenvectors and eigenvalues of L) is called the spectrum of the graph. It can be seen that

$$L(G) = \sum_i \lambda_i e_i e_i^T \tag{1.4}$$

The eigenvalues $\lambda_0, \ldots, \lambda_{n-1}$ represent the strength of projection of the matrix onto the basis elements. This may be viewed from a statistical point of view [31] where each $\lambda_i e_i e_i^T$ may be used to approximate $A(G)$ with approximation error inversely proportional to $1 - \lambda_i$. However, for a graph, those nodes which are best approximated by $\lambda_i e_i e_i^T$ in fact form a cluster of nodes. This is the basis for spectral clustering, a technique which uses the eigenvectors of L to perform clustering of a dataset or graph [26]. The first (smallest) non-zero eigenvalue and associated eigenvector are associated with the main clusters of data. Subsequent eigenvalues and eigenvectors can be associated with cluster splitting and also identification of smaller clusters [29]. Typically, there exists what is called a *spectral gap* in which for some k and j, $\lambda_k \ll \lambda_{k+1} \approx 1 \approx \lambda_{j-1} \ll \lambda_j$. That is, eigenvalues $\lambda_{k+1}, \ldots, \lambda_{j-1}$[1] are approximately equal to one and are likely to represent links in a graph which do not belong to any particular cluster. It is then usual to reduce the dimensionality of the data using an approximation based on the spectral decomposition. However, in this chapter we are interested in representing the global structure of a graph (e.g., we are interested in the presence or absence of many small clusters), which is essentially the spread of clustering across the graph. This information is contained in all the eigenvalues of the spectral decomposition.

[1] i.e., the eigenvalues at the center of the spectrum.

Let $x = (x_0, \ldots, x_{n-1})$ be a vector. From (1.3) we see that

$$xLx^T = \sum_{uv \in E} (x_u/\sqrt{d_u} - x_v/\sqrt{d_v})^2 \tag{1.5}$$

Now, the eigenvalues cannot be large because from (1.5) we obtain

$$\begin{aligned} xLx^T &\leq \sum_{uv \in E} (x_u/\sqrt{d_u} - x_v/\sqrt{d_v})^2 \\ &\qquad + (x_u/\sqrt{d_u} + x_v/\sqrt{d_v})^2 \\ &= 2\sum_u x_u^2 = 2xx^T \end{aligned} \tag{1.6}$$

and so $\lambda_i = e_i L e_i^T \leq 2$. What is more, the mean of the eigenvalues is 1 because

$$\sum_i \lambda_i = \mathrm{tr}(L) = n \tag{1.7}$$

by (1.3), where $\mathrm{tr}(L)$ is the $trace$ of L.

To summarize: the eigenvalues of L lie in the range 0 to 2 (the smallest being 0), i.e., $0 = \lambda_0 \leq \ldots \leq \lambda_{n-1} \leq 2$, and their mean is 1.

The distribution of the n numbers $\lambda_0, \ldots, \lambda_{n-1}$ contains useful information about the network, as will be seen. In turn, information about this distribution is given by its moments in the statistical sense, where the N^{th} moment is $1/n \sum_i (1-\lambda_i)^N$. These moments have a direct physical interpretation in terms of the network, as follows. Writing B for the matrix $D^{-1/2}AD^{-1/2}$, so that $L = I - B$, then by (1.3) the entries of B are given by

$$(D^{-1/2}AD^{-1/2})_{i,j} = \frac{A_{i,j}}{\sqrt{d_i}\sqrt{d_j}} \tag{1.8}$$

Now the numbers $1-\lambda_i$ are the eigenvalues of $B = I - L$, and so $\sum_i (1-\lambda_i)^N$ is just $\mathrm{tr}(B^N)$. Writing $b_{i,j}$ for the (i,j)-th entry of B, the (i,j)-th entry of B^N is the sum of all products $b_{i_0,i_1} b_{i_1,i_2} \ldots b_{i_{N-1} i_N}$ where $i_0 = i$ and $i_N = j$. But $b_{i,j}$, as given by (1.8), is zero unless nodes i and j are adjacent. So we define an N-cycle in G to be a sequence of vertices $u_1 u_2 \ldots u_N$ with u_i adjacent to u_{i+1} for $i = 1, \ldots, N-1$ and with u_N adjacent to u_1. (Thus, for example, a triangle in G with vertices set $\{a, b, c\}$ gives rise to six 3-cycles abc, acb, bca, bac, cab and cba. Note that, in general, an N-cycle might have repeated vertices.) We now have

$$\sum_i (1-\lambda_i)^N = \mathrm{tr}(B^N) = \sum_C \frac{1}{d_{u_1} d_{u_2} \ldots d_{u_N}} \tag{1.9}$$

the sum being over all N-cycles $C = u_1 u_2 \ldots u_N$ in G. Therefore, $\sum_i (1-\lambda_i)^N$ counts the number of N-cycles, normalized by the degree of each node in the cycle.

The number of N-cycles is related to various graph properties. The number of 2-cycles is just (twice) the number of edges and the number of 3-cycles is (six times)

the number of triangles. Hence $\sum_i (1-\lambda)^3$ is related to the clustering coefficient, as discussed below. An important graph property is the number of 4-cycles. A graph which has the minimum number of 4-cycles, for a graph of its density, is quasi-random, i.e., it shares many of the properties of random graphs, including, typically, high connectivity, low diameter, having edges distributed uniformly through the graph, and so on. This statement is made precise in [34] and [7]. For regular graphs, (1.9) shows that the sum $\sum_i (1-\lambda)^4$ is directly to the number of 4-cycles. In general, the sum counts the 4-cycles with weights: for the relationship between the sum and the quasi-randomness of the graph in the non-regular case, see the more detailed discussion in [8, Chapter 5]. The right hand side of (1.9) can also be seen in terms of random walks. A random walk starting at a vertex with degree d_u will choose an edge with probability $1/d_u$ and at the next vertex, say v, choose an edge with probability $1/d_v$ and so on. Thus the probability of starting and ending randomly at a vertex after N steps is the sum of the probabilities of all N-cycles that start and end at that vertex. In other words exactly the right hand side of (1.9). As pointed out in [35], random walks are an integral part of the Internet AS structure.

The left hand side of Equation (1.9) provides an interesting insight into graph structure. The right hand side is the sum of normalized N-cycles whereas the left hand side involves the spectral decomposition. We note in particular that the spectral gap is diminished because eigenvalues close to one are given a very low weighting compared to eigenvalues far from one. This is important as the eigenvalues in the spectral gap typically represent links in the network that do not belong to any specific cluster and are not therefore important parts of the larger structure of the network.

Next, we consider the well-known clustering coefficient. It should be noted that there is little connection between the clustering coefficient, and cluster identification, referred to above. The clustering coefficient, $\gamma(G)$, is defined as the average number of triangles divided by the total number of possible triangles

$$\gamma(G) = 1/n \sum_i \frac{T_i}{d_i(d_i-1)/2}, d_i \geq 2 \tag{1.10}$$

where T_i is the number of triangles for node i and d_i is the degree of node i. Now consider a specific triangle between nodes a, b and c. For the clustering coefficient, noting that the triangle will be considered three times, once from each node, the contribution to the average is

$$\frac{1}{d_a(d_a-1)/2} + \frac{1}{d_b(d_b-1)/2} + \frac{1}{d_c(d_c-1)/2} \tag{1.11}$$

However, for the weighted spectrum (with $N=3$), this particular triangle gives rise to six 3-cycles and contributes

$$\frac{6}{d_a d_b d_c} \tag{1.12}$$

So, it can be seen that the clustering coefficient normalizes each triangle according to the total number of possible triangles while the weighted spectrum (with $N=3$) instead normalizes using a product of the degrees. Thus, the two metrics can be

considered to be similar but not equal. Indeed, it should be noted that the clustering coefficient is in fact not a metric in the strict sense. While two networks can have the same clustering coefficient they may differ significantly in structure. In contrast, the elements of $\sum_i (1-\lambda)^3$ will only agree if two networks are isomorphic.

We now formally define the *weighted spectrum* as the normalized sum of N-cycles as

$$W(G,N) = \sum_i (1-\lambda_i)^N \tag{1.13}$$

However, calculating the eigenvalues of a large (even sparse) matrix is computationally expensive. In addition, the aim here is to represent the *global* structure of a graph and so precise estimates of *all* the eigenvalue values are not required. Thus, the distribution[2] of eigenvalues is sufficient. In this chapter the distribution of eigenvalues $f(\lambda = k)$ is estimated using pivoting and Sylvester's Law of Inertia to compute the number of eigenvalues that fall in a given interval. To estimate the distribution we use K equally spaced bins.[3] A measure of the graph can then be constructed by considering the distribution of the eigenvalues as

$$\omega(G,N) = \sum_{k \in K} (1-k)^N f(\lambda = k) \tag{1.14}$$

where the elements of $\omega(G,N)$ form the *weighted spectral distribution*:

$$WSD : G \rightarrow \Re^{|K|}\{k \in K : ((1-k)^N f(\lambda = k))\} \tag{1.15}$$

In addition, a metric can then be constructed from $\omega(G)$ for comparing two graphs, G_1 and G_2, as

$$\Im(G_1, G_2, N) = \sum_{k \in K} (1-k)^N (f_1(\lambda = k) - f_2(\lambda = k))^2 \tag{1.16}$$

where f_1 and f_2 are the eigenvalue distributions of G_1 and G_2 and the distribution of eigenvalues is estimated in the set K of bins $\in [0, 2]$. Equation (1.16) satisfies all the properties of a metric [14].

We next wish to test if the WSD for graphs generated by the same underlying process vary significantly (to show that the WSD is stable). To do this, we generate a set of graphs that have very similar structure and test to see if their WSDs are also similar. The results of an empirical test are shown in Figure 1.1. This plot was created by generating 50 topologies using the AB [1] generator with the (fixed) optimum parameters determined in §1.7, but with different initial conditions.[4] For each run the spectral and weighted spectral distributions are recorded yielding 50×50 bin values which are then used to estimate standard deviations. As the underlying model (i.e. the AB generator) is the same for each run, the *structure* might be expected

[2]The eigenvalues of a given graph are deterministic and so *distribution* here is not meant in a statistical sense.

[3]K can be increased depending on the granularity required.

[4]We found similar results for other parameters and topology generators.

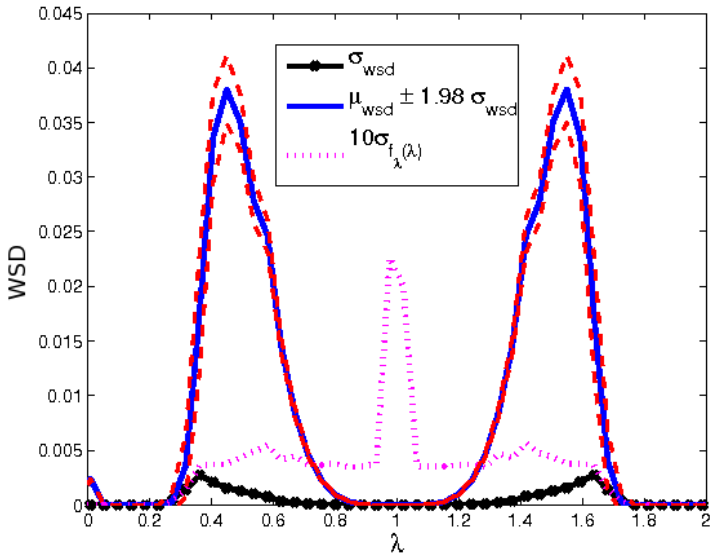


Figure 1.1 Mean and standard deviations for WSD and (unweighted) spectrum for the AB model over 50 simulations.

to remain the same and so a "structural metric" should be insensitive to random initial conditions. As can be seen the standard deviation[5] of the (unweighted) spectrum, $\sigma_{f_\lambda}(\lambda)$, is significantly higher at the center of the spectrum. However, for the WSD, the standard deviation, σ_{wsd}, peaks at the same point as the WSD; the noise in the spectral gap has been suppressed. The evidence suggests that the WSD successfully filters out the noise around1 in the middle region and highlights the important parts of the signal.

1.4 A SIMPLE WORKED EXAMPLE

After the fairly theoretical previous section, we aim at giving the reader a better intuition behind the WSD with a simple example. Figure 1.2 shows a small network, called G_1, with 7 nodes and 8 links. As can be seen there are 2 cycles of length 3 in this network and one of length 4. We will take $N = 3$ in this example for convenience and without loss of generality. The random walk probabilities are labeled in Figure 1.2. For example, node 3 has a degree of 5 resulting in a probability of 1/5$^{\text{th}}$

[5]Multiplied by a factor of ten for clarity.

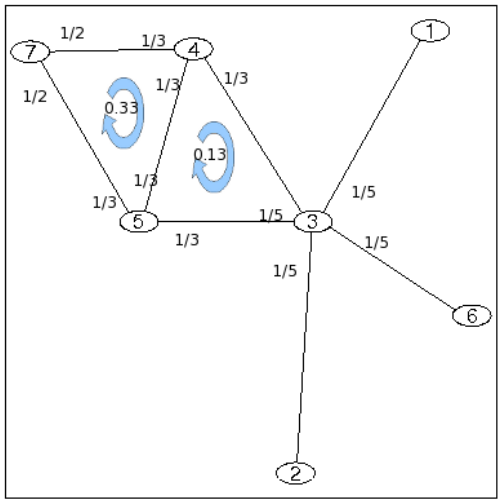


Figure 1.2 A simple example network G_1.

Table 1.1 Eigenvalues, WSD and dominant nodes of example network.

e_7	Eigenvector	λ	$1-\lambda$	$(1-\lambda)^3$	Dominant nodes
0.2500	1	1.8615	-0.8615	-0.6394	3,1,2,6
0.2500	2	1.3942	-0.3942	-0.0612	7,4,5
0.5590	3	1.3333	-0.3333	-0.0370	4,5
0.4330	4	1.0000	0.0000	0.0000	6,2
0.4330	5	1.0000	0.0000	0.0000	1,2,6
0.2500	6	0.4110	0.5890	0.2043	7,3
0.3536	7	0.0000	1.0000	1.0000	3,4,5,7
	$\sum_{i=1}^{7}(1-\lambda_i)^3$			0.4667	

for each edge. The total probability of taking a random walk around each 3-cycle is: $6 \times 1/2 \times 1/3 \times 1/3 = 0.33$, also shown.[6]

Figure 1.3 shows a 3-D plot of the absolute value (for clarity) of the eigenvectors of the normalized Laplacian. The corresponding eigenvalues are shown in Table 1.1.

[6]The six comes from the fact that the random walk can start in one of three nodes and go in one of two directions. It can be viewed in our case as really just a nuisance scaling factor.

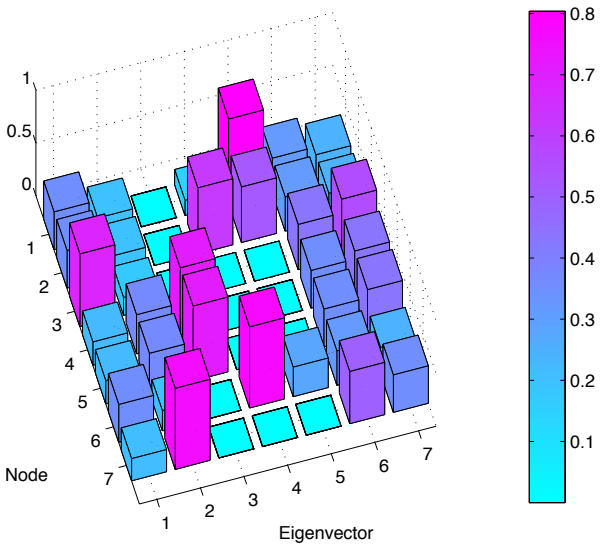


Figure 1.3 Eigenvectors of the simple example network.

As is well known, the eigenvectors of the normalised Laplacian perform a partitioning of the nodes in a graph. In this example nodes 4 and 5 are grouped into eigenvector 3, nodes 1,2 and 6 into eigenvectors 4 and 5, node 7 into eigenvector 2 and node 3 into eigenvector 1 (Figure 1.3). Note that for each partition the nodes in the partition are the same; i.e. we could swap the labels between nodes 4 and 5 and the network would not change (i.e. an isomorphism). Eigenvector and eigenvalue 7, e_7 and $\lambda_7 = 0$, are special and partitions all the nodes in the network with the most central nodes having the highest coefficients (see Table 1.1, column 1). In general the number of eigenvalues that are zero is equal to the number of components, arguably the most important structural property in a graph. This graph contains 1 connected component and so has a single zero eigenvalue (λ_7). Note that the highest possible weighting in the WSD is given at zero (i.e. 1 = 1-0); the number of components in the graph.

Note that the sum of the eigenvalues taken to the power of N is indeed the same as the sum of the probabilities of taking N random walk cycles in the graph. This is shown in Table 1.1, last row, $\sum_{i=1}^{7}(1-\lambda_i)^3 = 0.4667$ which can be easily verified by adding the cycle probabilities from Figure 1.3 ($0.3333 + 0.1333 = 0.467$). What is interesting is how this sum is constructed. In Table 1.1 the main contributions to the sum are from eigenvalues 1,2,3 and 6 (we ignore eigenvalue 7 as it merely reflects that the graph is connected) which are dominated by the nodes which form the cycles; 3, 4, 5 and 7.

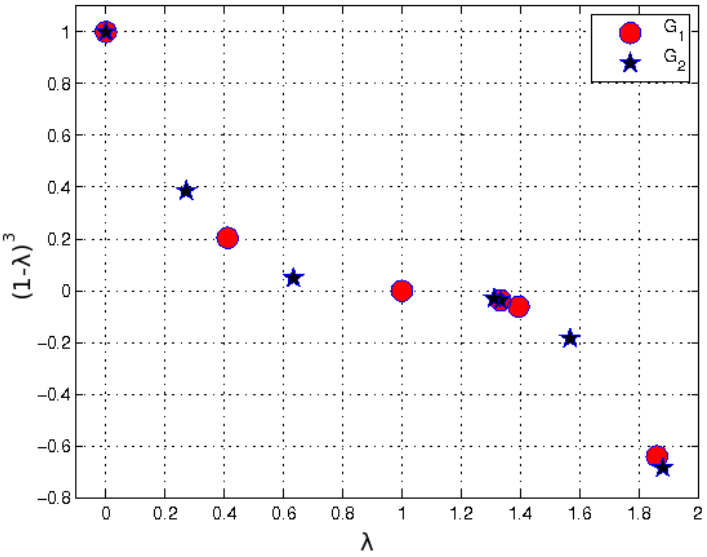


Figure 1.4 WSD of the example network.

However, this does not mean that the information provided by the WSD is confined to N-cycles in the graph. For example in Figure 1.5 we take the edge linking nodes 1 and 3 and rewire it so that 1 and 6 are now connected. Note that while the right cycle is still in place its probabilities have now changed, as the degree of node 3 is now 4. The corresponding eigenvalues have also changed as seen in Figure 1.4. [7]

In conclusion, the WSD can roughly be seen as an amalgamation of *local* views (i.e. walks of length N) taken from all the nodes. As $(1-\lambda_i) \leq 1\ \forall i$, $(1-\lambda_i)^N$ will suppress the smaller eigenvalues more and more as N increases [8]. We consider 3 and 4 to be suitable values of N for the current application: $N = 3$ is related to the well-known and understood clustering coefficient; and $N = 4$ as a 4-cycle represents two routes (i.e., minimal redundancy) between two nodes. For other applications, other values of N may be of interest. Also note that in section 1.3 we propose using the *distribution* of the eigenvalues for large networks; unfortunatly it is not instructive to talk about a distribution for a small number of eigenvalues (7 in this example).

[7]Note that if we had used the adjacency matrix instead of the normalised Laplacian the re-wiring would have no effect on the sum of the eigenvalues.

[8]This is closely related to the settling times in Markov chains which are often expressed in terms of the largest non-trivial eigenvalue. It differs in that the Walk Laplacian and not the normalised Laplacian is used.

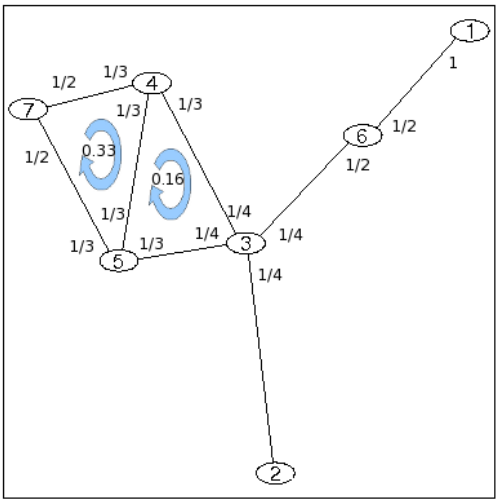


Figure 1.5 The second example network, G_2.

1.5 THE INTERNET AUTONOMOUS SYSTEM TOPOLOGY

The Internet's AS topology is a widely studied representation of the Internet at a particular scale. An AS represents a single network which can apply its own operational and peering policy. An Internet Service Provider (ISP) may use 1 or more ASes. The Internet contains over 30,000 ASes, each in a set of relationships with its neighbors, who are either its customers, providers or peers. In the Internet core there is a full mesh formed between the ASes of the various tier-1 ISPs. However, at the edge there are a huge number of smaller ISPs and customer networks which connect through upstream providers and local public exchange points. These smaller ISPs and customer networks may have only one upstream provider, or may have many for resilience and performance reasons. In addition, the Internet constantly evolves: new networks are added, old ones disappear and existing ones grow and merge.

Links between ASes depend on business relationships which can and do change, sometimes rapidly, making any interpretation of the Internet as a *static* structure inaccurate. This rich and dynamic structure makes it difficult to provide either a single, representative topological model, or a single graph metric that captures all characteristics of any topology. However, such a metric would make it possible to generate realistic synthetic topologies improving the accuracy of Internet-wide protocol simulations, and perhaps enabling the prediction of the future evolution of the Internet's topology.

Many attempts to capture one or more characteristics have been made, resulting in several topology generators which each synthesize Internet-like topologies using different models and parameters. Unfortunately, validating these models is an *ad hoc* matter that typically means matching several topological measures in the hope that this will ensure a matching structure. Users often select default parameters for these models based on specific datasets measured at particular times, which no longer represent the current Internet. However, as noted previously, these measures cannot be used to estimate the optimum parameters for a model given a target topology.

1.5.1 Characterization

Over the past several years many topological metrics have been proposed for quantitatively characterizing topological properties of networks. In this section we present a large set of topological metrics that will be used to measure a *distance* in graph space,[9] i.e., how topologically distant two graphs are from each other. These metrics are computed for both synthetic and measured AS topologies. When choosing our metrics we considered both those used by the topology generator designers and those used more widely in the graph theory literature. Taken individually, these metrics focus on different topological aspects, but when considered together they reveal a more complete picture of the observed AS topologies.

We specifically chose not to use the three metrics of Tangmunarunkit *et al.* [33] for two reasons. First, computation of resilience and distortion are both NP-complete, requiring the use of heuristics. In contrast, all our metrics are straightforward to compute directly. Second, although accurate reproduction of degree-based metrics is well-supported by current topology generators, our hypothesis is that local interconnectivity has been poorly understood, and so we add several metrics that focus on exactly this, e.g., assortativity, clustering, and centrality.

AS topologies are modeled as graphs $G = (V, E)$ with a collection of nodes V and a collection of links E that connect a pair of nodes. The number of nodes and links in a graph is then equal to, respectively, $N = |V|$ and $M = |E|$.

Degree. The degree k of a node is the number of links adjacent to it. The *average node degree* $\bar{k}$ is defined as $\bar{k} = 2M/N$. The *node degree distribution* $P(k)$ is the probability that a randomly selected node has a given degree k. The node degree distribution is defined as $P(k) = n(k)/N$, where $n(k)$ is the number of nodes of degree k. The *joint degree distribution* (JDD) $P(k, k')$ is the probability that a randomly selected pair of connected nodes have degrees k and k'. A summary measure of the joint degree distribution is the average neighbor degree of nodes with a given degree k, and is defined as follows $k_{nn}(k) = \sum_{k'=1}^{k_{max}} k' P(k'|k)$. The maximum possible $k_{nn}(k)$ value is $N - 1$ for a maximally connected network, i.e. a complete graph. Hence, we represent the JDD by the normalized value $k_{nn}(k)/(N-1)$ [23] and refer to it as *average neighbor connectivity*.

[9] In [16] we present an even larger set of measures.

Assortativity. Assortativity is a measure of the likelihood of connection of nodes of similar degrees [28]. This is usually expressed by means of the *assortativity coefficient* r: assortative networks have $r > 0$ (disassortative have $r < 0$ resp.) and tend to have nodes that are connected to nodes with similar (dissimilar resp.) degree.

Clustering. Given node i with k_i links, these links could be involved in at most $k_i(k_i-1)/2$ triangles (e.g., nodes $a \rightarrow b \rightarrow c \rightarrow a$ form a triangle). The greater the number of triangles, the greater the clustering of this node. The clustering coefficient $\gamma(G)$ is defined as the average number of triangles divided by the total number of possible triangles: $\gamma(G) = 1/N \sum_i \frac{T_i}{k_i(k_i-1)/2}, k_i \geq 2$ where T_i is the number of triangles of node i and k_i is its degree. We use the distribution of *clustering coefficients* $C(k)$, which in fact is the distribution of the terms $\frac{T_i}{k_i(k_i-1)/2}$ in the overall summation. This definition of the clustering coefficient gives the same weight to each triangle in the network, irrespective of the distribution of the node degrees.

Rich-Club. The *rich-club coefficient* $\phi(\rho)$ is the ratio of the number of links in the component induced by the ρ largest-degree nodes to the maximum possible links $\rho(\rho-1)/2$, where $\rho = 1...N$ are the first ρ nodes ordered by their degree ranks in a network of size N nodes and ρ is normalized by the total number of nodes N [9, 41]. In this way, the node rank ρ denotes the position of a node on this ordered list.

Shortest Path. The shortest path length distribution $P(h)$ is the probability distribution of two nodes being at minimum distance h hops from each other. From the shortest path length distribution the average node distance in a connected network is derived as $\overline{h} = \sum_{h=1}^{h_{\max}} hP(h)$, where $h_{\max}$ is the longest shortest path between any pair of nodes. $h_{\max}$ is also referred to as the diameter of a network.

Centrality. Betweenness centrality is a measure of the number of shortest paths passing through a node or a link. The *node betweenness* for a node v is $B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through a node v [19]. The average node betweenness is the average value of the node betweenness over all nodes.

Closeness. Another measure of the centrality of a node within a network is its *closeness*. The closeness of a node is the reciprocal of the sum of shortest paths from this node to all other reachable nodes in a graph.

Coreness. The l-core of a network (sometimes known as the k-core) is the maximal component in which each node has at least degree l. In other words, the l-core is the component of a network obtained by recursively removing all nodes of degree less than l. A node has coreness l if it belongs to the l-core but not to the $(l+1)$-core. Hence, the l-core is the collection of all nodes having coreness l. The core of a network is the l-core such that the $(l+1)$-core is empty [4].

Clique. A clique in a network is a set of pairwise adjacent nodes, i.e., a component which forms a complete graph. The *top clique size*, also known as the graph clique number, is the number of nodes in the largest clique in a network [38].

Spectrum. It has recently been observed that eigenvalues are closely related to almost all critical network characteristics [8]. For example, Tangmunarunkit *et al.* [33] classified network resilience as a measure of network robustness subject to link failures, resulting in a minimum balanced cut size of a network. Spectral graph theory enables study of network partitioning using graph eigenvalues [8]. In this chapter we focus on the spectrum of the *normalized Laplacian matrix*, where all eigenvalues lie between 0 and 2, allowing easy comparison of networks of different sizes. We use the normalized graph's spectrum for tuning the parameters of topology generators.

1.5.2 Generation

In this section we present a number of topology generators, each having their own set of parameters. We also present an example of an Internet AS topology dataset which we use as a litmus test for the parameter tuning exercise.

There are many models available that claim to describe the Internet AS topology. Several of these are embodied in tools built by the community for generating simulated topologies. In this section we describe the particular models whose output we compare in this chapter. The first are produced from the Waxman model [36], derived from the Erdös-Rényi random graphs [12], where the probability of two nodes being connected is proportional to the Euclidean distance between them. The second come from the Barabasi and Albert (BA) [3] model, following measurements of various power laws in degree distributions and rank exponents by Faloutsos *et al.* [13]. These incorporate common beliefs about preferential attachment and incremental growth. The third are from the Generalized Linear Preference model [5] which additionally model clustering coefficients. Finally, Inet [37] and PFP [41] focus on alternative characteristics of AS topology: the existence of a meshed core, and the phenomenon of preferential attachment respectively. Each model focuses only on particular metrics and parameters, and has only been compared with selected AS topology observations [39, 33, 37].

Waxman. The Waxman model of random graphs is based on a probability model for interconnecting nodes of the topology given by:

$$P(u, v) = \alpha e^{-d/(\beta L)} \tag{1.17}$$

where $0 < \alpha, \beta \leq 1$, d is the Euclidean distance between two nodes u and v, and L is the network diameter, i.e., the largest distance between two nodes. Note that d and L are not parameters for the Waxman model. The Internet is known not to be a random network but we include the Waxman model as a baseline for comparison purposes.

BA. The BA [1] model was inspired by the idea of preferentially attaching new nodes to existing well-connected nodes, leading to the incremental growth of nodes and the

links between them. Starting with a network of m_0 isolated nodes, $m \leq m_0$ new links are added with probability p. One end of each link is attached to a random node, while the other end is attached to a node selected by preferring the more popular, i.e., well-connected, nodes with probability

$$\Pi(k_i) = \frac{k_i + 1}{\sum_j k_j + 1} \tag{1.18}$$

where k_j is the degree of node j, with probability q, m links are rewired and new nodes are added with probability $1 - p - q$. A new node m has m new links that, with probability $\Pi(k_i)$, are connected to nodes i already present in the system. We use the BRITE [25] implementation of this model in this chapter.

GLP. Our third model is the Generalized Linear Preference model (GLP) [5]. It focuses on matching characteristic path length and clustering coefficients. It uses a probabilistic method for adding nodes and links recursively while preserving selected power law properties. In the GLP model, when starting with m_0 links, the probability of adding new links is defined as p where $p \in [0, 1]$. Let $\Pi(d_i)$ be the probability of choosing node i. For each end of each link, node i is chosen with probability $\Pi(d_i)$ defined as:

$$\Pi(d_i) = (d_i - \beta) / \sum_j (d_j - \beta) \tag{1.19}$$

where $\beta \in (-\infty, 1)$ is a tunable parameter indicating the preference of nodes to connect to existing popular nodes. We use the BRITE implementation of this model in this chapter.

Inet. Inet [37] produces random networks using a preferential linear weight for the connection probability of nodes after modeling the core of the generated topology as a full mesh network. Inet sets the minimum number of nodes at 3037, the number of ASes on the Internet at the time of Inet's development. By default, the fraction of degree 1 nodes α is set to 0.3, based on measurements from Routeviews[10] and NLANR[11] BGP table data in 2002.

PFP. In the Positive Feedback Preference (PFP) model [41], the AS topology of the Internet is considered to grow by interactive probabilistic addition of new nodes and links. It uses a nonlinear preferential attachment probability when choosing older nodes for the interactive growth of the network, inserting edges between existing and newly added nodes. As the PFP generator does not have any user-tunable parameters we include it only in the last part of §1.7 for completeness.

[10] `http://www.routeviews.org/`
[11] `http://www.nlanr.net/`

1.5.3 Observations

The AS topology can be inferred from two main sources of data, BGP and traceroutes, both of which suffer from measurement artifacts. BGP data is inherently incomplete no matter how many vantage points are used for collection. In particular, even if BGP updates are combined from multiple vantage points, many peering and sibling relationships are not observed [15]. Traceroute data misses alternative paths since routers may have multiple interfaces which are not easily identified, and multi-hop paths may be hidden by tunnelling via Multi-Protocol Label Switching (MPLS). In addition, mapping traceroute data to AS numbers is often inaccurate [24].

Chinese. The first dataset is a traceroute measurement of the Chinese AS Topology collected from servers within China in May 2005. It reports 84 ASs, representing a small subgraph of the Internet. Zhou *et al.* [40] claim that the Chinese AS graph exhibits all the major topology characteristics of the global AS graph. The presence of this dataset enables us to compare the AS topology models at smaller scales. Further, this dataset is believed to be nearly complete, i.e., it contains very little measurement bias and accurately represents the AS topology of that region of the Internet. Thus, although it is rather small, we have included it as a valuable comparison point in our studies.

Skitter. The second dataset comes from the CAIDA Skitter project.[12] By running traceroutes towards a large range of IP addresses and subsequently mapping the prefixes to AS numbers using RouteViews BGP data, CAIDA computes an observation of the AS topology. For our study we use the graphs from March 2004 to match those used by Mahadevan *et al.* [23]. This AS topology reports 9,204 unique ASs.

RouteViews. The third dataset we use is derived from the RouteViews BGP data. This is collected both as static snapshots of the BGP routing tables and dynamic BGP data in the form of BGP update and withdrawal messages. We use the topologies provided by Mahadevan *et al.* [23] from both the static and dynamic BGP data from March 2004. The dataset is produced by filtering AS sets and private ASs and merging the 31 daily graphs into one. This dataset reports 17,446 unique ASs across 43 vantage points in the Internet.

UCLA. The fourth dataset comes from the Internet topology collection[13] maintained by Oliviera *et al.* [30]. These topologies are updated daily using BGP routing tables and updates from RouteViews, RIPE,[14] Abilene[15] and LookingGlass servers. We use a snapshot of this dataset from November 2007, computed using a time window on

[12] `http://www.caida.org/tools/measurement/Skitter/`
[13] `http://irl.cs.ucla.edu/topology/`
[14] `http://www.ripe.net/db/irr.html`
[15] `http://abilene.internet2.edu/`

the last-seen timestamps to discard ASs which have not been seen for more than 6 months. The resulting dataset reports 28,899 unique ASs.

1.6 COMPARING TOPOLOGY GENERATORS

Most past comparisons of topology generators have been limited to the average node degree, the node degree distribution and the joint degree distribution. The rationale for choosing these metrics is that if those properties are closely reproduced, then the value of other metrics will also be closely reproduced [22].

In this section we show that current topology generators are able to match first and second order properties well, i.e., average node degree and node degree distribution, but fail to match many other topological metrics. We also discuss the importance of various metrics in our analysis.[16]

1.6.1 Methodology

For each generator we specify the required number of nodes and generate 10 topologies of that size to provide confidence intervals for the metrics. We then compute the metrics introduced in §1.5 on both the generated and observed AS topologies. All topologies studied in this chapter are undirected, preventing us from considering peering policies and provider-customer relationships. This limitation is forced upon us by the design of the generators as they do not take such policies into account.

Each topology generator uses several parameters, all of which could be tuned to best fit a particular size of topology. However, there are two problems with attempting this tuning. First, doing so requires selecting an appropriate goodness-of-fit measure, of which there are many as noted in §1.5. Second, in any case tuning parameters to a particular dataset is of questionable merit since, as we argued in §1.2, each dataset is but a sample of reality, having many biases and inaccuracies. Typically, topology generator parameters are tuned to match the number of links in the synthetic and measured networks for a given number of nodes. However we found this to be infeasible as generating graphs with equal numbers of links from a random model and a power-law model gives completely different outputs. For space reasons we deal with this particular issue elsewhere [18]; in this chapter we simply use the default values embedded within each generator.

1.6.2 Topological metrics

In this section we discuss the results for each metric separately and analyze the reasons for differences between the observed and the generated topologies.

Table 1.2 displays the values of various metrics (columns) computed for different topologies (rows). Blocks of rows correspond to a single observed topology and the

[16]We present an extended set of metrics in [16] which further support our claims; we restrict ourselves here to only the most significant results.

Table 1.2 Comparison of AS level dataset with synthetic topologies.

Topology	Links	Avg. deg.	Max. degree	Top clique size	Max. betweenness	Max. coreness	Assort. coef.	Clust. coef.	Max. closeness
Chinese	*211*	*5.02*	*38*	*2*	*1,324*	*5*	*-0.32*	*0.188*	*<0.01*
Waxman	252	6	18	**2**	404	4	0.039	0.117	**0.506**
BA	165	3.93	19	3	**1,096**	2	-0.096	0.073	0.515
GLP	151	3.6	44	3	2,391	**5**	-0.257	**0.119**	0.643
PFP	**250**	**5.95**	**37**	10	849	9	**-0.38**	0.309	0.638
Skitter	*28,959*	*6.3*	*2,070*	*16*	*10,210,533*	*28*	*-0.23*	*0.026*	*<0.01*
Waxman	**27,612**	**6**	33	0	474,673	4	0.205	0.002	**0.264**
BA	18,405	4	190	0	5,918,226	2	-0.05	0.001	0.315
GLP	16,744	3.64	**2,411**	2	34,853,544	5	-0.089	0.003	0.496
INET	18,504	4.02	1,683	3	15,037,631	7	-0.195	0.004	0.514
PFP	27,611	**6**	3,000	**16**	**13,355,194**	**24**	**-0.244**	**0.012**	0.588
RouteViews	*40,805*	*4.7*	*2,498*	*9*	*30,171,051*	*28*	*-0.19*	*0.02*	*<0.01*
Waxman	52,336	6	35	0	1,185,687	4	0.205	0.001	**0.25**
BA	34,889	4	392	3	33,178,669	2	-0.04	0.001	0.33
GLP	31,391	3.6	4,226	4	127,547,256	6	-0.08	0.002	0.48
INET	**43,343**	**4.97**	**2,828**	**6**	**31,267,607**	14	-0.258	0.006	0.522
PFP	52,338	6	4,593	23	39,037,735	**30**	**-0.252**	**0.009**	0.564
UCLA	*116,275*	*8.05*	*4,393*	*10*	*76,882,795*	*73*	*-0.165*	*0.05*	*0.32*
Waxman	86,697	6	40	0	3,384,114	4	0.213	<0.001	0.246
BA	57,795	4	347	0	52,023,288	2	-003	<0.001	**0.3**
GLP	52,456	3.63	7391	2	371,651,147	6	-0.08	<0.001	0.486
INET	**91,052**	**6.3**	**6,537**	**12**	**88,052,316**	38	-0.3	**0.01**	0.55
PFP	86,696	6	8076	26	123,490,676	**40**	**-0.218**	**0.01**	0.57

generated topologies with the same number of nodes as the observed topology. Bold numbers represent nearest match of a metric value to that for the relevant observed topology. Rows in each block are ordered with the observed topology first, followed by the generated topologies from oldest to newest generator. For synthetic topologies, the value of the metrics is averaged over the 10 generated instances. Note that Inet requires the number of nodes to be greater than 3037 and hence cannot be compared to the Chinese topology.

We observe a small but measurable improvement from older to newer generators in some metrics such as maximum degree, maximum coreness, and assortativity coefficient. This suggests that topology generators have successively improved at matching particular properties of the observed topologies. However, the number of links in the generated topologies may differ considerably from the observed topology due to the assumptions made by the generators. The Waxman and BA generators fail to capture the maximum degree, the top clique size, maximum betweenness and coreness. Those two generators are too simplistic in the assumptions they make about the connectivity of the graphs to generate realistic AS topologies. Waxman relies on a random graph model which cannot capture the clique between tier-1 ASes nor the heavy tail of the node degree distribution. BA tries to reproduce the power-law node degrees with its preferential attachment model but fails to reach the maximum node degree, as it only adds edges between new nodes and not between existing ones. Hence, neither of these two models is able to create the highly-connected core of tier-1 ASes. PFP and Inet manage to come closer to the values of the metrics of the observed topologies. For Inet this is because it assumes that 30% of the nodes are fully meshed (at the core), whereas for PFP its rich-club connectivity model allows to add edges between existing nodes.

Node degree distribution. In Figure 1.6 we show the CCDF of the node degree for all topologies on a log-log scale. We observe that the Chinese topology does not exhibit power law scaling due to its limited size, whereas all the larger AS topologies do exhibit power-law scaling of node degrees. The Waxman generator completely fails to capture this behavior as it is based on a random graph model, but recent topology generators do capture this power law behavior of the node degrees quite well, as observed in [5]. In the case of the RouteViews and UCLA datasets, Inet and PFP outperform other topology generators. Note that the more complete UCLA dataset has a slightly concave shape in contrast to RouteViews where the degree distribution displays strict power law scaling. In summary, more recent generation models reproduce node degree distributions well as expected since this has been a primary focus in the literature.

Average neighbor connectivity. Neighbor connectivity has been far less studied than node degree, although it is very important to match local interconnection among a node's neighbors when reproducing the topological structure of the Internet [23]. Figure 1.7 shows the CCDF of the average neighbor degrees for all topologies. We observe that Waxman, BA and GLP underestimate the local interconnection structures

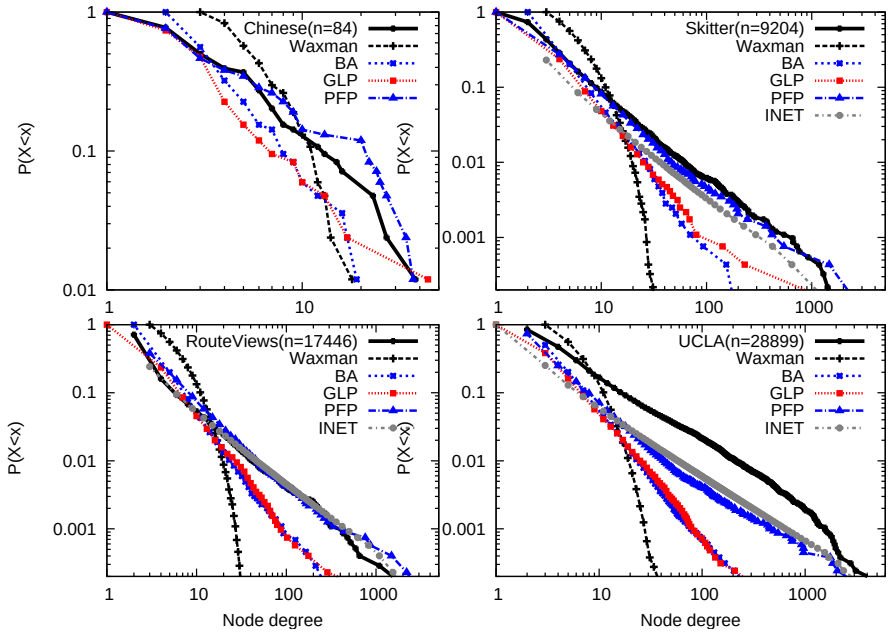


Figure 1.6 Comparison of node degree CCDFs.

around nodes. BA and GLP typically generate graphs with far fewer links than the observed topologies so they underestimate neighbor degrees on average.

For the larger observed topologies, i.e., RouteViews and UCLA, PFP and Inet typically overestimate the neighbor connectivity, as they both place a large number of inter-AS links in the core. In addition, the shapes of the neighbor connectivity CCDF differ for the larger topologies: Inet and PFP have two regimes, one for highly connected nodes (those with larger neighbor connectivity), and another for low-degree nodes. On the other hand, observed topologies have a smooth region for the high-degree nodes followed by a rather stable region caused by similar degree nodes. We observe that the highest degree nodes in the UCLA topology have very high values of neighbor connectivity. This is consistent with the belief that tier-1 providers are densely meshed.

Clustering coefficients. Like the average neighbor connectivity, the clustering coefficient gives information about local connectivity of the nodes. It is important to reproduce clustering due to its impact on the local robustness in the graph: nodes with higher local clustering have increased local path diversity [23].

Figure 1.8 displays the clustering coefficients of all nodes in the topologies. Error bars indicate 95% confidence intervals around the mean values of the 10 topologies from each generator. We observe that Waxman and BA significantly underestimate

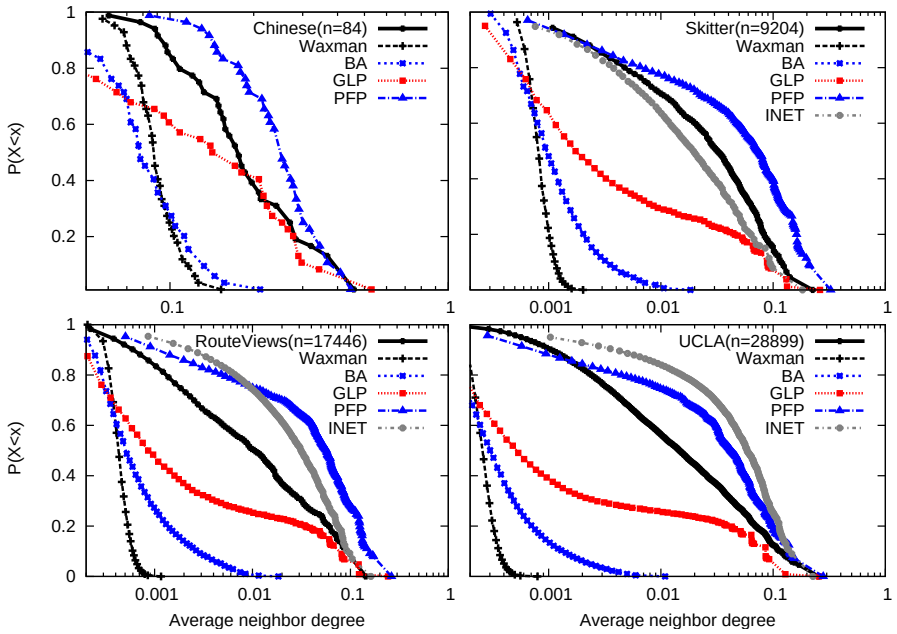


Figure 1.7 Comparison of average neighbor connectivity CCDFs.

clustering, consistent with their simplistic way of connecting nodes. GLP approximates the clustering of the Chinese topology quite well but fails in the case of the larger observed topologies. PFP and Inet capture clustering reasonably well compared to the other topology generators. However, Inet does not reproduce the tail of the distribution well due to the randomness factor in its model for edge addition once the core is fully meshed.

We also observe that for medium degree nodes, clustering coefficients display rather high variability which increases with the size of the observed topologies. This behavior seems to be a property of the observed AS topology of the Internet.

In summary, all topology generators fail to properly capture clustering, typically underestimating local connectivity. Only Inet for the UCLA topology overestimates connectivity of low-degree nodes while still underestimating it for high-degree nodes. Current topology generators do not seem to adequately model of local node connectivity.

Rich-club connectivity. Rich-club connectivity gives information about how well-connected nodes of high degree are among themselves. Figure 1.9 makes it clear that the cores of the observed topologies are very close to a full mesh, with values close to 1 on the left of the graphs. The error bars again indicate the 95% confidence intervals around the mean values of the different instances of the generated topologies.

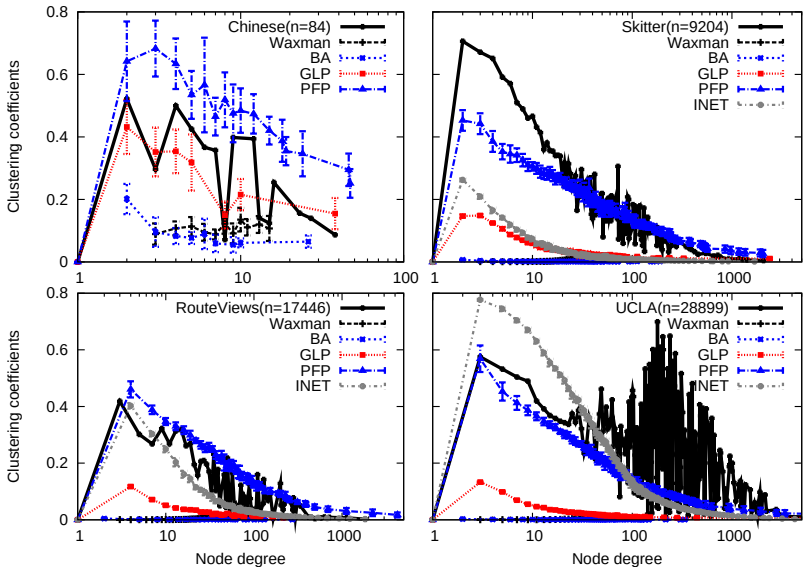


Figure 1.8 Comparison of clustering coefficients.

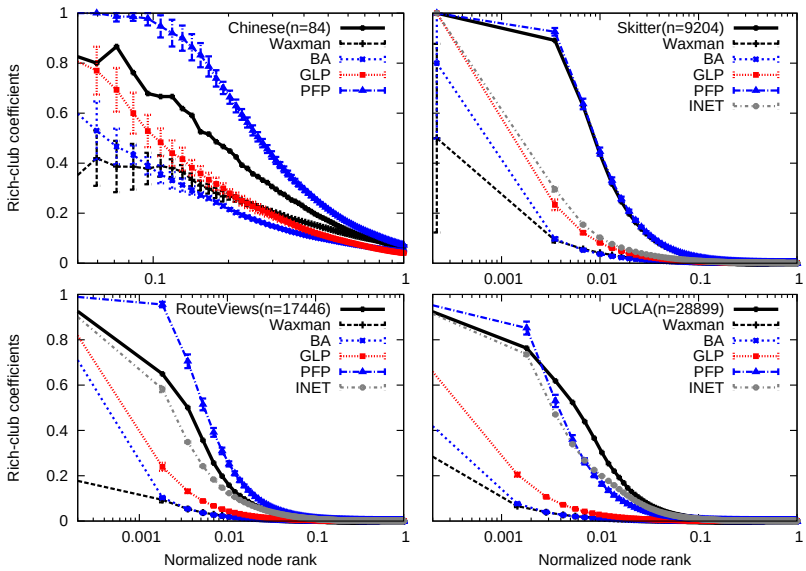


Figure 1.9 Comparison of rich-club connectivity coefficients

Waxman and BA perform poorly for this metric in general. Only PFP and Inet generate topologies with a dense enough core compared to the observed topologies. Given the emphasis that PFP gives to the rich-club connectivity, it overestimates it in the case of the Chinese and RouteViews topologies. Inet performs well due to its emphasis on a highly connected core, especially for larger topologies where data has been collected across multiple peering points.

In summary, most topology generators underestimate the importance of rich-club connectivity of the AS topology. PFP is the only topology generator that emphasizes the importance of the dense core of the AS topology.

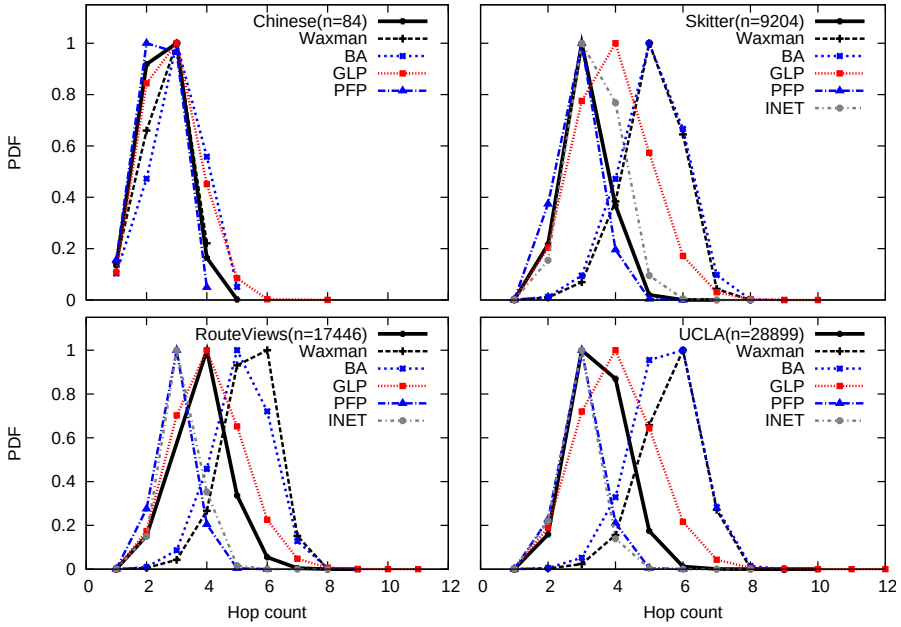


Figure 1.10 Comparison of shortest path distributions (number of hops).

Shortest path distributions. Figure 1.10 displays the distributions of shortest path length. Apart from BA, most topology generators approximate the shortest path length distribution of the Chinese graph quite well due to its small size. For the other topologies, PFP and Inet generally underestimate the path length distribution while Waxman and BA overestimate it. Particular generators seem to capture the path length distribution for particular topologies well: PFP matches Skitter's well and GLP is close for Routeviews. Inet and PFP both do a better job for UCLA than for RouteViews but both still underestimate the distribution.

In summary, shortest path length is not well captured by any topology generator. As shortest path length is related to local connectivity, failing to capture local connectivity is likely to lead to such a behavior.

Spectrum. The spectrum of the normalized Laplacian matrix is a powerful tool for characterizing properties of a graph. If two graphs have the same spectrum, they have the same topological structure.

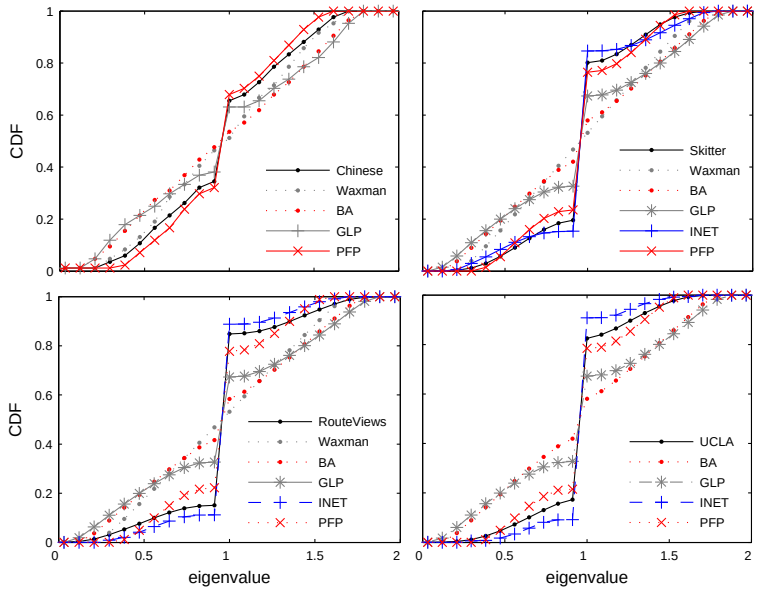


Figure 1.11 Comparison of cumulative distributions of eigenvalues (from normalized Laplacian).

Figure 1.11 displays the CDF of the eigenvalues computed from the normalized Laplacian matrix of each topology.

As with other topological metrics, Inet and PFP perform best. The difference between the topology generators is most easily observed around the eigenvalues equal to 1. These eigenvalues play a special role as they indicate repeated duplications of topological patterns within the network. By duplication, we mean different nodes having the same set of neighbors giving their induced subgraphs the same structure. Through repeated duplication, one can create networks with high multiplicity of eigenvalue 1 [2]. Further, if a network is bipartite, i.e., it consists of two connected parts with no links between nodes of the same part, then its spectrum will be symmetric about 1. This phenomenon can also arise through repeated structure duplication.

We observe that the spectra have a high degree of symmetry around the eigenvalue 1, and so the observed AS topologies appear close in spectral terms to a bipartite graph. In the AS topology, many ASes share a similar set of upstream ASes without being directly connected to each other. Inet and PFP are good examples of topology generators where this strategy is implemented. Note that the simple preferential attachment model of BA does not reproduce the eigenvalues around 1 very well. In the simple BA model, new nodes connect randomly to a given number of existing nodes, favoring connections to high degree nodes. In the Internet in contrast, although small ASes may tend to connect to large upstream providers, they might not connect preferentially to the largest ones, connecting instead to national or regional providers. In summary, these results provide further evidence that the interconnection structure of the AS topology is more complex than current models assume.

1.6.3 Discussion

Deviations between topology models and observations have been already studied in the literature. However, most works so far have focussed on particular topological metrics. Concentrating on particular topological metrics has lead to underestimate the mismatch between the properties of observed AS topologies and what current models produce. When comparing several models with several observed AS topologies as we do, we see that current topology models mostly try to capture some properties of one set of observations from the AS topology. For a topology model to claim to model the Internet's AS topology, we would expect that it tries to approach the properties of observed AS topologies in many respects, which is not the case today.

1.7 TUNING TOPOLOGY GENERATOR PARAMETERS

The aim of this section is to examine how well the topology generators match the Skitter topology for different values of their parameters. To facilitate this comparison, grids are constructed over the possible values of the parameter spaces and various cost functions are evaluated as follows:

1. A cost function measuring the matching between the number of links in Skitter and the generated topologies:

$$C_1(\theta) = (l_t(\theta) - l_{skitter})^2 \tag{1.20}$$

   where C_1 is the first cost function, θ are the model parameters (which differ for each topology generator), l_t is the number of links (which is a function of the parameters) and $l_{skitter}$ is the number of links in the Skitter dataset.

2. A cost function measuring the matching between the spectra of the Skitter network and of the generated topologies:

$$C_2(\theta) = \sum_i (P(\Lambda \leq \lambda_{t,i}) - P(\Lambda \leq \lambda_{skitter,i}))^2 \tag{1.21}$$

where $\lambda_{t,i}$ is the i^{th} eigenvalue for topology t.

3. A cost function measuring the matching of the weighted spectra:

$$C_3(\theta) = \sum_i ((w * P(\Lambda = \lambda_{t,i}) - w * P(\Lambda = \lambda_{skitter,i}))^2 \qquad (1.22)$$

where weight $w = (1 - i)^4$.

In addition to examining different parameter values across a grid, the optimum parameters with respect to $C_3(\theta)$ are estimated using the Nelder Meade simplex search algorithm [27, 11]. Note that the topologies generated by the topology generators are random in a statistical sense, due to differing random seeds for each run. Ten topologies are generated for each value of θ and the average spectral distribution is calculated. We found that the variance of the spectral distributions was sufficiently low to allow reasonable estimates of the minima in each case.

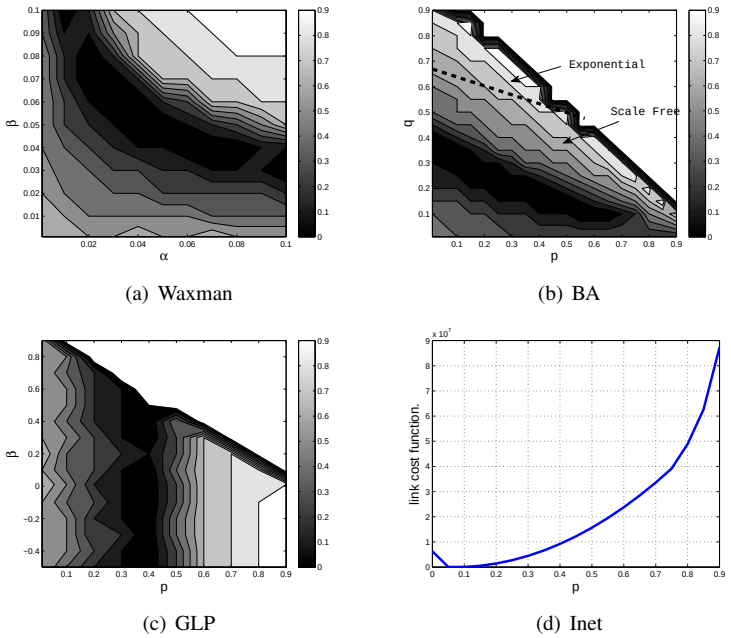
1.7.1 Link Densities

Figure 1.12 displays the value of the cost function $C_1(\theta)$ as a function of the topology generator parameters. On the upper and lower left graphs, the grayscale color indicates the value of the cost function. For Inet (lower right) there is only one parameter, p, so it is plotted as a curve in Figure 1.12(d). Figure 1.12 shows that a minimum exists for each topology in approximately the same regions as the default values of each generator.[17] For the BA generator + it is known that for values of p and q above the line shown in Figure 1.12(b), the topologies generated follow an exponential node degree distribution while those below follow a scale-free distribution. It is encouraging to note that the values of $C_1(\theta)$ are large in the exponential region and the minimum is in the scale-free region as the node degree distribution of the Internet is known to be approximately scale free [1]. Overall the results obtained by tuning the parameters based on $C_1(\theta)$ appear reasonable. For link density matching it is possible to obtain parameter values which match the link densities exactly. Indeed, there is a ridge of parameters for BA, GLP and Waxman for which the link densities can be matched. However, as noted in the introduction, there is no control over any other characteristic of the graph using this method.

1.7.2 Spectra PDF

Figure 1.13 shows the spectral PDF of the Skitter dataset and the four topology generators calculated at three parameters values in each grid (the parameter values are indicated in brackets in the legends). The aim is to illustrate how much the spectral PDFs change with the values of the parameters. The spectral PDFs of Waxman (Figure 1.13(a)) vary significantly for different values of α and β. Furthermore, none of the Waxman PDFs match well the spectral PDF of the Skitter graph. The BA PDFs

[17] Some of these default values are listed in Table 1.3.



(a) Waxman (b) BA

(c) GLP (d) Inet

Figure 1.12 Topology generator parameter grid for sum squared error from number of links.

vary to a lesser extent (Figure 1.13(b)) and appear to give a much better match than the Waxman model, especially around eigenvalue 1 ($\lambda = 1$). This better match of BA is not surprising as the Waxman model is not a good model for the Internet as noted in §1.6. GLP (Figure 1.13(c)) and Inet (Figure 1.13(d)) give similar results to BA, with a poor match outside eigenvalue 1. The better match of the BA model around eigenvalue 1 is interesting. As noted in §1.3 the regions away from eigenvalue 1 are far more important than the region around $\lambda = 1$. However, what is required is a technique that reveals the differences with distance from one as these are more important. Thus it would appear difficult to evaluate which model, or even which parameter, is better based on the PDFs alone. This point is now further explored by analysis of the grids calculated with respect to $C_2(\theta)$.

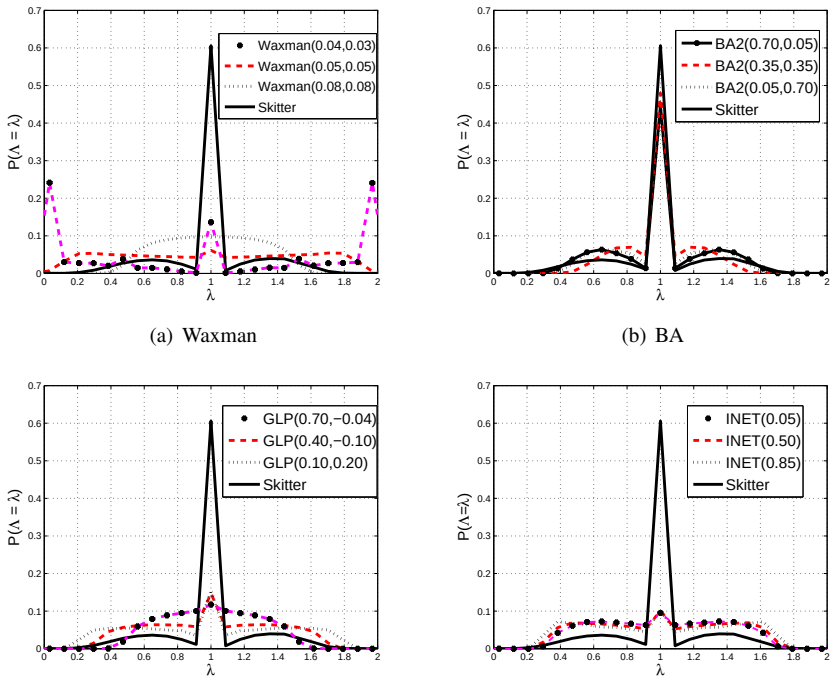
1.7.3 Limitations of Spectra CDF

Figure 1.14 shows the value of the second cost function $C_2(\theta)$ as a function of the topology generator parameters, in the same way as Figure 1.12. As can be seen in Figure 1.14, there are many islands corresponding to local minima, creating a rugged landscape. The variance in the PDFs referred to in this section is actually greater than any gradient that might exist in the grid. This means that it is not possible to estimate the minimum with respect to $C_2(\theta)$. Figure 1.14 shows that the spectrum on its own is not sufficient to identify the optimum parameters of any of the topology generators.

This is because each eigenvalue in $C_2(\theta)$ is weighted equally. As noted in §1.3, the eigenvalues close to 1 are more likely to be affected by the random seeds for each topology generator and are the source of the noise on the grid.

1.7.4 Weighted Spectra

The previous section illustrated the limitations of using the raw eigenvalues to find optimal topology generator parameters to match the Skitter topology. Figure 1.15 shows a plot of the weighted spectra of the same topologies as those shown on Figure 1.13. As can be seen the results are quite different from those shown in Figure 1.13. The Waxman weighted spectra still shows a bad fit with respect to the Skitter data (mainly around 0 and 2) compared to the other generators. The other generators (BA, GLP and Inet) now show that they are capable of matching the weighted spectra of the Skitter topology, especially around the point of greatest weight ($\lambda = 0.4$ or 1.6). The difference between the weighted spectra around 1 is no longer of importance (in contrast to Figure 1.13), reflecting that the weights here approach zero as we approach eigenvalue 1. In the next section the optimum values and the resulting weighted spectra will be compared.



(a) Waxman (b) BA

(c) GLP (d) Inet

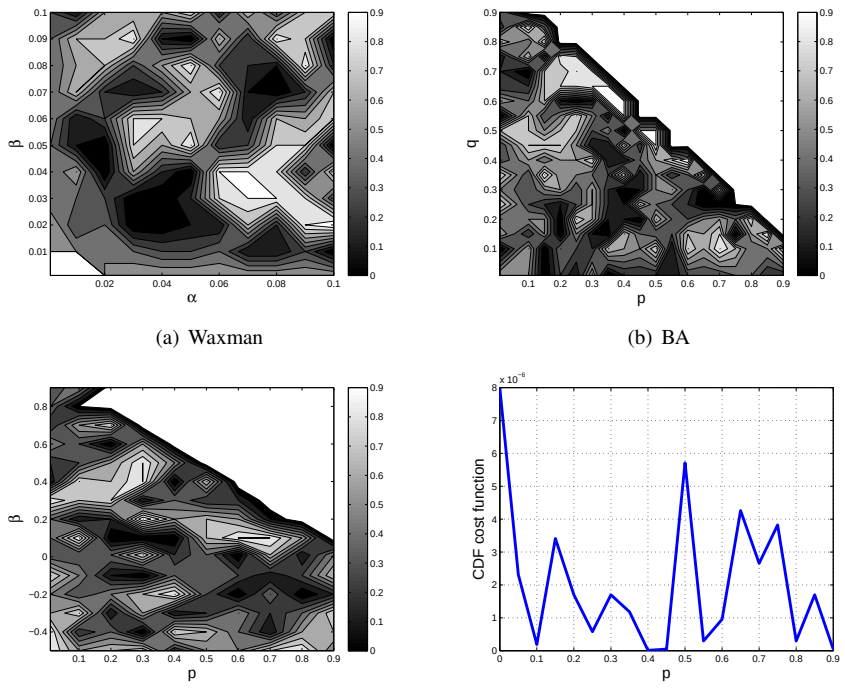
Figure 1.13 PDF of Spectra

1.7.5 Weighted Spectra Comparison

Figure 1.16 shows the grids associated with $C_3(\theta)$. As can be seen the grids show that there is a region with a minima in each case and in addition, comparing Figure 1.16 and Figure 1.12 it can be seen that these minima lie in a region close to those for $C_1(\theta)$. However, it should be noted that the weighted spectra will try to fit more than just the number of links in a topology. This demonstrates the inherent trade-off. Also of note is that the region of interest for the BA model lies inside the region of scale-free behavior as shown in Figure 1.16(b).

1.8 GENERATING TOPOLOGIES WITH OPTIMUM PARAMETERS

Table 1.3 displays the optimum values for the topology generators for generating networks that are close to the Skitter graph. In addition, we give the values for $C_3(\theta)$, which show that PFP gives the closest fit followed by BA, GLP, Waxman and finally Inet. While these results are mostly expected, the ranking of Inet as the worst topology generator is surprising. We have also listed some of the default parameters used in certain generators such as BRITE [25]. While many of the optimized parameters are



(a) Waxman (b) BA

(c) GLP (d) Inet

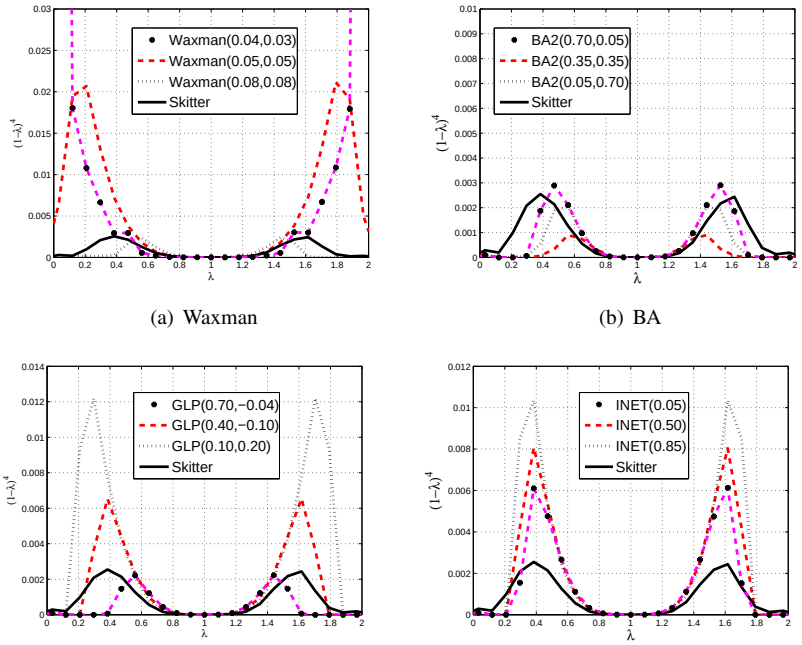
Figure 1.14 Parameter grid for sum of absolute differences of spectra CDFs.

close to the default values, which is encouraging, it should be noted that the default parameters are for a *typical* graph and are not selected for any particular situation. Thus a direct comparison is meaningless.

Table 1.3 Optimum parameter values for matching Skitter topology.

Generator	Optimum and Default Parameter Values		$C_3(\theta)$	$\overline{C_3}(\theta)$
Waxman	$\alpha = 0.08$ (def. 0.15)	$\beta = 0.08$ (def. -0.2)	0.0026	0.0797
BA	$p = 0.2865$ (def. 0.6)	$q = 0.3145$ (def. 0.3)	0.0014	0.0300
GLP	$p = 0.5972$ (def. 0.45)	$\beta = 0.1004$ (def. 0.64)	0.0021	0.0446
Inet	$\alpha = 0.1013$ (def. 0.3)	–	0.0064	0.0150
PFP	–	–	0.0014	0.0371

Figure 1.17(a) shows the weighted spectra for each of the topology generators and inspection of this figure goes some way to explaining the discrepancy in the results. As can be seen the main peak in the weighted spectra for the Skitter data occurs at a value of $\lambda = 0.4$. The Waxman generator peak occurs at $\lambda = 0.6$ which is closer to 1 demonstrating the greater amount of random structure in the Waxman topologies. However, for the Inet generator the peak occurs at the correct point ($\lambda = 0.4$) but the

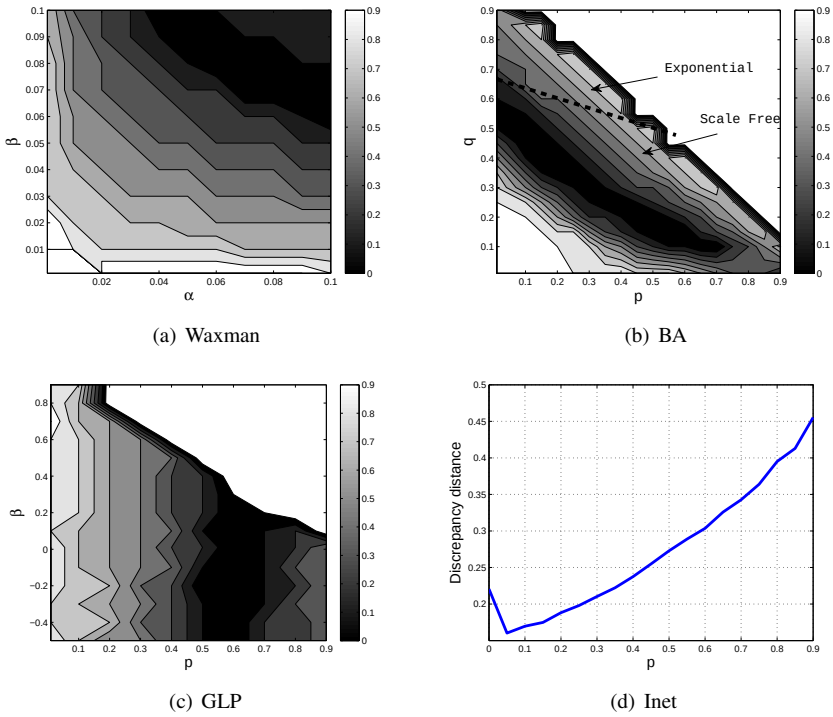


(a) Waxman (b) BA

(c) GLP (d) Inet

Figure 1.15 Weighted spectra grid for generator parameters.

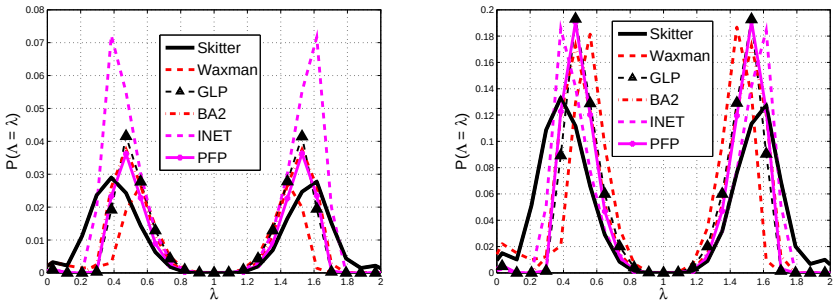
weighted power at this point is far greater than in the Skitter topology. By normalizing the weighted spectrum this point becomes clear:



(a) Waxman (b) BA

(c) GLP (d) Inet

Figure 1.16 Grid of sum squared error of weighted spectra for topology generators



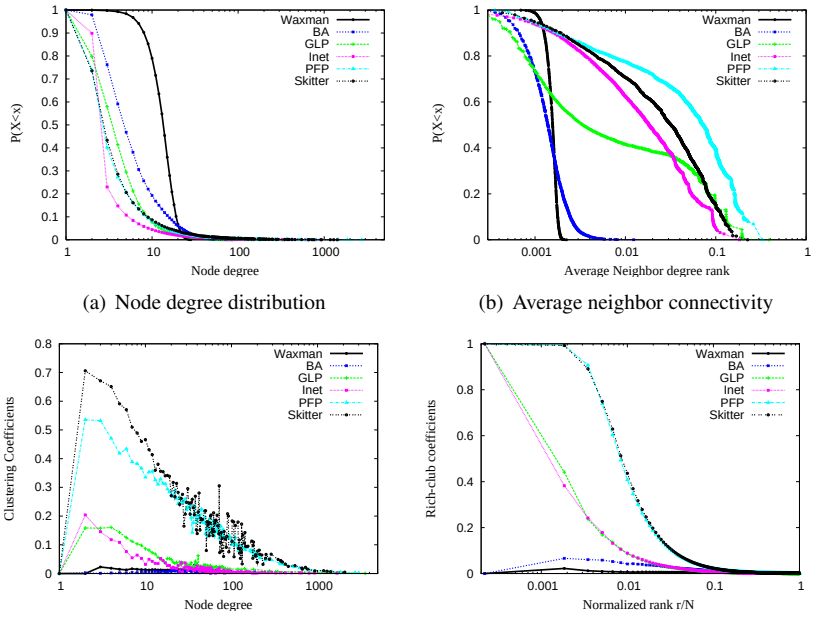
(a) Weighted spectra (b) Normalized weighted spectra

Figure 1.17 Comparison of the weighted spectra.

$$\overline{C_3}(\theta) = \sum_i \frac{((w_i * P(\Lambda = \lambda_{t,i}))}{\sum_i ((w_i * P(\Lambda = \lambda_{t,i}))} - \frac{((w_i * P(\Lambda = \lambda_{skitter}))}{\sum_i ((w_i * P(\Lambda = \lambda_{skitter}))} \quad (1.23)$$

Using the normalized weighted spectrum the results in Figure 1.17(b) show that Inet is the best match for the Skitter data while the Waxman model still performs worse than the other models. Further research is required before stating which version of C_3 is superior.

Figure 1.18 shows a comparison of the optimized topologies with respect to four typical network metrics: the node degree distribution, the average neighbor connectivity, the clustering coefficient and the rich-club connectivity [41]. As can be seen PFP gives the best match for these metrics in agreement with our proposed metric $C_3(\theta)$. The performance of the other topologies is mixed showing that while one topology is able to match one metric it fails to match another. For example, the GLP generator achieves a reasonable match for the node degree distribution but fails to match the average neighbor connectivity. It is interesting to note that BA does not match the rich club connectivity which is not evident in our metric.



(a) Node degree distribution

(b) Average neighbor connectivity

(c) Clustering coefficients

(d) Rich-Club connectivity

Figure 1.18 Comparison of topology generators and Skitter topology.

1.9 INTERNET TOPOLOGY EVOLUTION

The WSD produces a mapping from $\Re^{M\times M} \longmapsto \Re^{|K|}$, where $|K| = 71$ bins are used in the examples in this section. However, a 71 dimensional space is still too large to effectively visualise clustering across graphs. In this section, we introduce *Multi-Dimensional Scaling* (MDS), a technique mapping the WSD into a lower dimension.

Specifically, given C different graphs we seek a mapping from their WSD's into an l dimensional space: $\Re^{C\times|K|} \longmapsto \Re^{C\times l}$ where $l << |K|$. Typically $l = 2$ or 3 makes visual inspection most straightforward. Note that the methods used are parameter-free and so a *natural* clustering of the data is sought, as opposed to a supervised method which applies a mapping learned from training data.

Multi-Dimensional Scaling (MDS) [10] is a technique mapping *distances* between objects into a reduced dimensional space. An intuitive example involves taking the distance matrix commonly shown in the bottom corner of many road maps and using it to reconstruct the map itself. The technique uses *distance* between the graphs here defined in terms of the metric introduced in Equation 1.16, $\Im(G_1, G_2, N)$. First, a dissimilarity matrix, R, is constructed as:

$$R_{(i,j)} = \begin{cases} \Im(G_i, G_j, N) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{1.24}$$

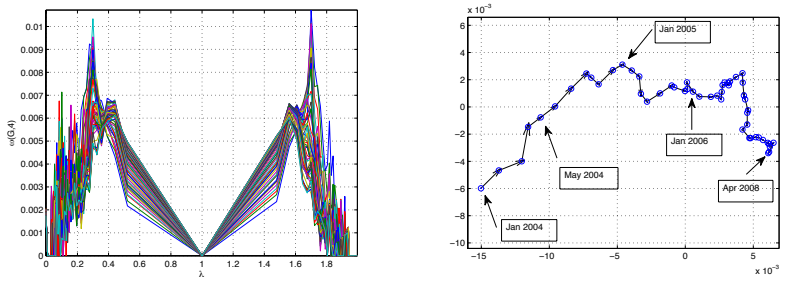
The goal of MDS is to find a set of vectors $Z_1, Z_2, ...Z_{|K|}$ that incrementally approximate the distance in the dissimilarity matrix. Specifically, we wish to minimise the distance between the projected vectors and the orignal data as:

$$C = \min_{Z_1,Z_2,...Z_{|K|}} \sum_{i<j} (\|Z_i - Z_j\| - R_{(i,j)})^2 \tag{1.25}$$

where C is the cost function to be minimised. We then perform the minimisation using numerical optimisation based on the eigenvector decomposition of R [32]. Typically, the first and second vectors, Z_1 and Z_2, are sufficient to allow visualisation of clustering within the data.

Figure 1.19 shows the evolution of the Internet AS topology over time, as observed in the UCLA dataset described in §1.5.3. It is difficult to discern any consistent evolution from the raw WSD plots in Figure 1.19(a). However, applying the MDS to reduce the dimensionality from 71 to 2 results in Figure 1.19(b), in which each point represents the projection of a computed WSD for a given topology, i.e., the WSD computed for a given month's observations in the UCLA dataset. Note that the axes are dimensionless: it is not the particular values that are important but the separation of points computed.

Interestingly, plotting with an arrow joining consecutive points, i.e., an arrow connects the points for datasets 1 and 2, another connects points for datasets 2 and 3, &c., shows that the evolution of the WSD for the topology appears to be consistent over time: it represents the "structural walk" of the Internet AS topology observed by the UCLA data. The lack of clustering of points around a centre suggests the structure of the Internet is evolving in some way. This evolution is very difficult



(a) Raw WSDs. (b) WSDs after applying MDS.

Figure 1.19 Structural evolution of the Internet via raw WSD and WSD with MDS applied.

to see by directly comparing the WSD lines but can easily be observed using this multi-dimensional scaling technique. This is much more straightforward than the current alternative approach which would involve using a complex set of topological measures to distinguish the different graphs [20]. The reason for this actual evolution is better examined in a different domain; for the interested reader we recommend reading [17]. Here the aim is merely to show that MDS used in conjunction with WSD can be used to track the structural changes in a network.

1.10 CONCLUSIONS

Comparison of graph structures is a frequently encountered problem across many scientific areas. To perform a meaningful comparison requires the definition of a cost-function that encodes those features of each graph considered important. While the spectrum of a graph encodes a graph's features, the raw spectrum contains too much information to be useful on its own. In this chapter we have introduced a new metric, the *weighted spectral distribution*, that improves on the raw graph spectrum by discounting those eigenvalues believed to be less significant and noisy, while emphasizing the contribution of those believed to be important and information-rich.

We then showed the use of this cost-function to optimize the selection of parameter values for the subject of Internet topology generation. The cost-function defined by the weighted graph spectrum was shown to lead to parameter choices that are appropriate in the context of the particular problem domain: Internet topology generation. In particular, we showed that the parameter choices so made are close to the default values and, in for one particular graph-generator (BA), fall within the expected region. In addition, as the metric is formed through summation, it is possible to go further and identify the particular eigenvalues that are responsible for significant differences. Although it is currently difficult to assign specific features to specific eigenvalues, we hope that this will also become a feature of the *weighted spectral distribution* in the future. Finally we briefly demonstrated a technique for projecting the raw WSD distributions into a lower dimensional space. This makes comparison of different dis-

tributions straightforward, as shown by the clear evolution of the Internet's topology viewed through the UCLA dataset.

REFERENCES

1. Reka Albert and Albert-Laszlo Barabasi. Topology of evolving networks: local events and universality. *Physical Review Letters*, 85, 2000.
2. A. Banerjee and J. Jost. Spectral plot properties: Towards a qualitative classification of networks. In *European Conference on Complex Systems*, October 2007.
3. A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, (1999).
4. Michael Baur, Ulrik Brandes, Marco Gaertler, and Dorothea Wagner. Drawing the AS graph in 2.5 dimensions. In János Pach, editor, *Graph Drawing, New York, 2004*, pages 43–48. Springer, 2004.
5. T. Bu and D. Towsley. On distinguishing between Internet power law topology generators. In *Proceedings of IEEE Infocom 2002*, June 2002.
6. H. Bunke. Graph matching: Theoretical foundations, algorithms, and applications. In *Proc. of the International Conference on Vision Interface*, pages 82–88, May 2000.
7. F. R. K. Chung, R. L. Graham, and R. M. Wilson. Quasi-random graphs. *Combinatorica*, 9(4):345–362, 1989.
8. Fan R. K. Chung. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics)*. American Mathematical Society, (1997).
9. V. Colizza, A. Flammini, M.A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115, 2006.
10. T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994.
11. J.E. Dennis and D.J. Woods. Optimization in microcomputers: The nelder-meade simplex algorithm. In A. Wouk, editor, *New Computing Environments: Microcomputers in Large-Scale Computing*, pages 116–122. SIAM, (1987).
12. P. Erdös and A. Rényi. On random graphs. In *Mathematical Institute Hungarian Academy, 196*, London, (1985).
13. Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of ACM SIGCOMM 1999*, (1999).
14. Damien Fay, Hamed Haddadi, Andrew G. Thomason, Andrew W. Moore, Richard Mortier, Almerima Jamakovic, Steve Uhlig, and Miguel Rio. Weighted Spectral Distribution for Internet Topology Analysis: Theory and Applications. *IEEE/ACM Transactions on Networking (ToN)*, 18(1):164–176, 2010.
15. Anja Feldmann, Olaf Maennel, Z. Morley Mao, Arthur Berger, and Bruce Maggs. Locating Internet routing instabilities. In *Proceedings of ACM SIGCOMM 2004*, 2004.
16. H. Haddadi, D. Fay, A. Jamakovic, O. Maennel, A. W. Moore, R. Mortier, M. Rio, and S. Uhlig. Beyond node degree: Evaluating AS topology models. Technical Report UCAM-CL-TR-725, University of Cambridge, Computer Laboratory, July 2008.

17. Hamed Haddadi, Damien Fay, Steve Uhlig, Andrew Moore, Richard Mortier, and Almerima Jamakovic. Mixing biases: Structural changes in the AS topology evolution. In *Proceedings of the 2nd Traffic Monitoring and Analysis (TMA) Workshop*, Zurich, Switzerland, April 2010.
18. Hamed Haddadi, Damien Fay, Steve Uhlig, Andrew Moore, Richard Mortier, Almerima Jamakovic, and Miguel Rio. Tuning topology generators using spectral distributions. In *Lecture Notes in Computer Science, Volume 5119, SPEC International Performance Evaluation Workshop*, Darmstadt, Germany, 2008. Springer.
19. P. Holme, B.J. Kim, C.N. Yoon, and S.K. Han. Attack vulnerability of complex networks. *Physical Review E*, 65(5):298–305, 2002.
20. Marios Iliofotou, Michalis Faloutsos, and Michael Mitzenmacher. Exploiting Dynamicity in Graph-based Traffic Analysis: Techniques and Applications. In *ACM CoNEXT*, 2009.
21. V. Kann. On the approximability of the maximum common subgraph problem. In *Proc. 9th Annual Symposium on Theoretical Aspects of Computer Science*, pages 377–388, 1992.
22. Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic topology analysis and generation using degree correlations. In *Proceedings of ACM SIGCOMM 2006*, pages 135–146, Pisa, Italy, 2006.
23. Priya Mahadevan, Dmitri Krioukov, Marina Fomenkov, Xenofontas Dimitropoulos, k c claffy, and Amin Vahdat. The Internet AS-level topology: three data sources and one definitive metric. *SIGCOMM Computer Communication Review*, 36(1):17–26, 2006.
24. Zhuoqing Morley Mao, Jennifer Rexford, Jia Wang, and Randy H. Katz. Towards an accurate AS-level traceroute tool. In *Proceedings of ACM SIGCOMM 2003*, pages 365–378, Karlsruhe, Germany, 2003.
25. Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. BRITE: an approach to universal topology generation. In *IEEE MASCOTS*, pages 346–353, Cincinnati, OH, USA, August (2001).
26. Boaz Nadler, Stephane Lafon, Ronald Coifman, and Ioannis Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Neural Information Processing Systems (NIPS)*, (2005).
27. J.A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, (1965).
28. M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):871–898, 2002.
29. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, (2002).
30. Ricardo Oliveira, Beichuan Zhang, and Lixia Zhang. Observing the Evolution of Internet AS Topology. In *Proceedings of ACM SIGCOMM 2007*, Kyoto, Japan, August 2007.
31. A.J. Seary and W.D. Richards. Spectral methods for analyzing and visualizing networks: an introduction. In *Dynamic Social Network Modeling and Analysis*, pages 209–228. National Academic Press, 2003.
32. G. A. F. Seber. *Multivariate Observations*. John Wiley & Sons, 1984.
33. Hongsuda Tangmunarunkit, Ramesh Govindan, Sugih Jamin, Scott Shenker, and Walter Willinger. Network topology generators: degree-based vs. structural. In *Proceedings of ACM SIGCOMM 2002*, pages 147–159, Pittsburgh, PA, 2002.

34. A. G. Thomason. Pseudo-random graphs. *Random Graphs '85, North-Holland Mathematical Study*, 144:307–331, 1987.

35. X. Wang and D. Loguinov. Wealth-based evolution model for the internet as-level topology. In *Proc. of IEEE INFOCOM*, April 2006.

36. Bernard M. Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications (JSAC)*, 6(9):1617–1622, December (1988).

37. Jared Winick and Sugih Jamin. Inet-3.0: Internet topology generator. Technical Report CSE-TR-456-02, University of Michigan Technical Report CSE-TR-456-02, (2002).

38. D. R. Wood. An algorithm for finding a maximum clique in a graph. *Operations Research Letters*, 21(7):211–217, January 1997.

39. Ellen W. Zegura, Kenneth L. Calvert, and Michael J. Donahoo. A quantitative comparison of graph-based models for Internet topology. *IEEE/ACM Transactions on Networking (TON)*, 5(6):770–783, (1997).

40. S. Zhou, G.-Q. Zhang, and G.-Q. Zhang. Chinese Internet AS-level topology. *IET Communications*, 1(2):209–214, April 2007.

41. Shi Zhou. Characterising and modelling the Internet topology, the rich-club phenomenon and the PFP model. *BT Technology Journal*, 24, (2006).