On the Importance of Local Connectivity for Internet Topology Models

Hamed Haddadi*, Damien Fay†, Almerima Jamakovic‡, Olaf Maennel ||,
Andrew W. Moore§, Richard Mortier¶, Steve Uhlig||
*Max Planck Institute for Software Systems (MPI-SWS)
†McGill University
‡TNO Netherlands
§University of Cambridge Computer Laboratory
¶Vipadia Ltd
||TU Berlin/Deutsche Telekom Laboratories

Abstract—**Existing models for Internet Autonomous System (AS) topology generation make structural assumptions about the AS graph. Those assumptions typically stem from beliefs about the true properties of the Internet, e.g. hierarchy and power-laws, which arise from incorrect interpretations of incomplete observations of the AS topology. In this paper we compare AS topology generation models with several observed AS topologies without making assumptions as to the relative importance of different topological characteristics. We find that although existing AS topology models capture degree-based properties well, they fail to capture the complexity of the local interconnection structure between ASes.**

We use a wide range of metrics including the *weighted spectral distribution* and make it available as toolbox[1]. We show that the shortcomings of existing models stem from underestimating the complexity of connectivity in the core due to incomplete understanding of collected data limitations, and narrow focus on particular aspects of the AS topology structure.

I. Introduction

For many years researchers have modeled the Internet's Autonomous System (AS) topology using graphs obtained via two main measurement techniques, i.e., BGP routing tables [1], [2] and traceroute maps [3]. The AS topology is an abstraction of the Internet commonly used to analyze its macro-level characteristics and to simulate the performance and scalability of new protocols and applications. Accurate simulation on Internet-scale topologies requires accurate AS topology generation models that match the observed topology across a wide range of metrics.

In this paper we evaluate existing AS topology generation models by comparing them with four available datasets that represent observed Internet AS topologies. A key principle underlying our work is to be agnostic about the topological properties of the Internet: we consciously avoid making assumptions as to the relative importance of the many topological properties. The main reason behind our agnosticism is the dynamic behavior of the Internet topology: it is constantly changing and so it is difficult to pick a particular metric as the most important when the fundamental nature of the underlying topology is evolving. In addition, observations of the AS topology suffer from two problems. First, a single set of observation points have only limited visibility of the topology. Second, each observation technique suffers from measurement artifacts, e.g., IP-to-AS number mapping and traceroute aliasing [4]. As a result, AS topology models make use of simplifying assumptions about the actual topology [5], [6]. One widely held assumption, based on biased observations, is that the AS topology has a hierarchical structure [7] and its node-degree distribution obeys a power-law [8].

In this comparison we rely on a wide set of commonly used topological measures[2], including a metric based on the graph spectrum (eigenvalues of the normalized Laplacian matrix) introduced by Fay *et al.* [10]. By using an extensive set of metrics we can observe differences in the topological properties of observed and synthetic AS topologies. We then go on to comparing the effects of using more measurement points for collecting topology data. This effort shows that an increase in number of measurement points increases the discovery of links between neighbors and hence the clustering features of the graph, while not greatly affecting its degree distribution.

This paper is structured as follows. Section II presents a set of available topology models. In Section III we present a set of observed AS topologies, collected using different methodologies from various locations. In Section IV, we present the results of our comparison and analyze the effect of adding measurement points in Section V. Finally, in Section VI we contrast our work with related work and in Section VII we conclude and discuss potential improvements in the field of AS topology modeling.

II. AS topology models

In this section we describe several models that try to reproduce properties of Internet AS topology datasets. Several of these models are embodied in topology generators [4].

Waxman: The Waxman model [11] derives from the Erdös-Rényi random graphs [12], where the probability of two nodes

[1] Available at http://www.cl.cam.ac.uk/research/srg/netos/masts/wsd.html/

[2] For a complete description of measures refer to [9]

being connected is proportional to the Euclidean distance between them. The probability of interconnecting nodes is $P(u,v) = \alpha\ e^{-d/(\beta L)}$, 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. We use the BRITE [13] implementation of this model, which ensures there are no disconnected components in the generated topology by re-wiring using iterative assignment of edges.

BA2: The Albert and Barabasi [14], the second model introduced by authors after [15] model was inspired by observations of various power laws in degree distributions and rank exponents by Faloutsos *et al.* [8]. The BA model is based on preferential attachment of new nodes to existing well-connected nodes and on the incremental growth of the number of nodes and the links between them. When a node i joins the network, the probability that it connects to an existing node j is $P(i,j) = \frac{d_j}{\sum_{k \in V} d_k}$, where d_j is the degree of node j, V is the set of nodes that have joined the network and $\sum_{k \in V} d_k$ is the sum of degrees of all nodes that previously joined the network [13].

GLP: The Generalized Linear Preference model (GLP) [5] focuses on matching characteristic path length and clustering coefficients. It probabilistically adds nodes and links while preserving selected power law properties.

Inet: Inet [16] 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. Inet sets the minimum number of nodes at 3037, the number of ASs in the Internet at the time of its development. It similarly sets the fraction of nodes having degree 1 to 0.3, based on measurements from Routeviews[3] and NLANR[4] BGP tables data in 2002.

PFP: The Positive Feedback Preference (PFP) model [17], assumes that the AS topology grows 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 nodes as well as the newly added ones.

III. AS TOPOLOGY 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 [18]. 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 [19].

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.* [20] 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[5]. 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.* [21]. 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.* [21] 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[6] maintained by Oliviera *et al.* [22]. These topologies are updated daily using BGP routing tables and updates from RouteViews, RIPE[7], Abilene[8] and LookingGlass servers. We use a snapshot of this dataset from November 2007, computed using a time window on 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.

IV. RESULTS AND DISCUSSION

Most past comparisons of topology generators have been limited to the average node degree, the node degree distribution and the joint degree distribution (see Section VI). 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 [6].

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 important topological metrics. These higher order statistics are critical for representiveness of the topologies [21]. We also discuss the importance of various metrics in our analysis[9].

[3] http://www.routeviews.org/

[4] http://www.nlanr.net/

[5] http://www.caida.org/tools/measurement/Skitter/

[6] http://irl.cs.ucla.edu/topology/

[7] http://www.ripe.net/db/irr.html

[8] http://abilene.internet2.edu/

[9] We present an extended set of metrics in [9] which further support our claims; we restrict ourselves here to only the most significant results here.

A. 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 [9] on both the generated and the observed AS topologies. All topologies studied in this paper 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. Second, tuning parameters to a particular dataset is of questionable merit since, as we argued in Section I, 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 in this paper we simply use the default values embedded within each generator by its designers and refer the reader to [23] for an analysis of the parameter tuning exercise.

B. 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 I displays the values of various metrics (columns) computed for different topologies (rows). Blocks of rows correspond to a single observed topology and the generated topologies with the same number of nodes as the observed topology. Rows in each block are ordered with the observed topology first, followed by the generated topologies from oldest to newest generator. Bold numbers represent nearest match of a metric value to that for the relevant observed topology. 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.

A small but measurable improvement is visible from older to newer generators in some metrics such as maximum degree, maximum coreness, and assortativity coefficient. Topology generators have successively improved at matching particular properties of the observed topologies. Notice the number of links in the generated topologies that differs 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 core 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 the Internet AS toplogy. 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.

1) Node degree distribution: Figure 1 displys the CCDF of the node degree for all topologies on a log-log scale. 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.

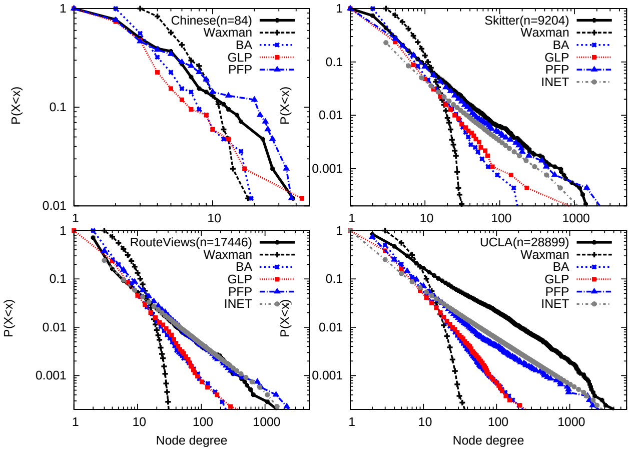

Fig. 1: Comparison of node degree CCDFs.

2) 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 [21]. Figure 2 shows the CCDF of the average neighbor degrees for all topologies. Waxman, BA and GLP underestimate the local interconnection structures 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

TABLE I: 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

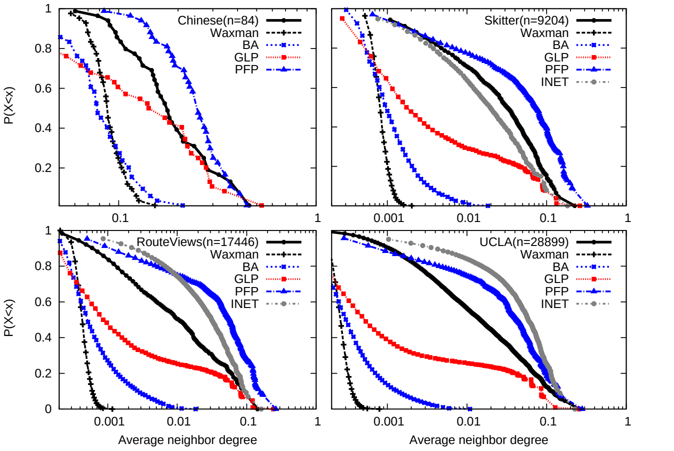

Fig. 2: Comparison of average neighbor connectivity CCDFs.

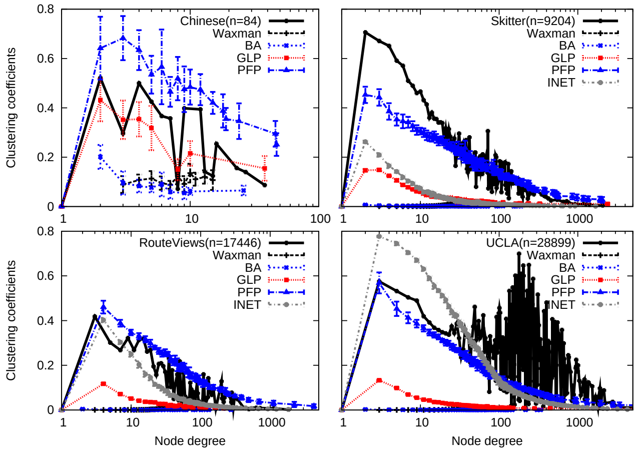

Fig. 3: Comparison of clustering coefficients.

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 another region caused by similar degree nodes. 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.

3) 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 [21].

Figure 3 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. Waxman and BA significantly underestimate 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 its random edge addition procedure once the core is fully meshed.

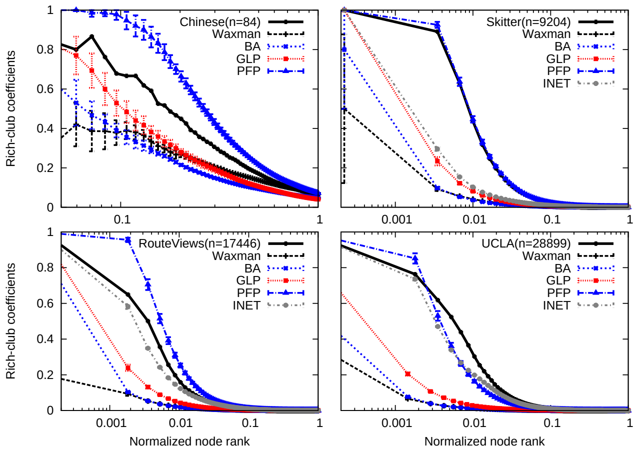

Fig. 4: Comparison of rich-club connectivity coefficients

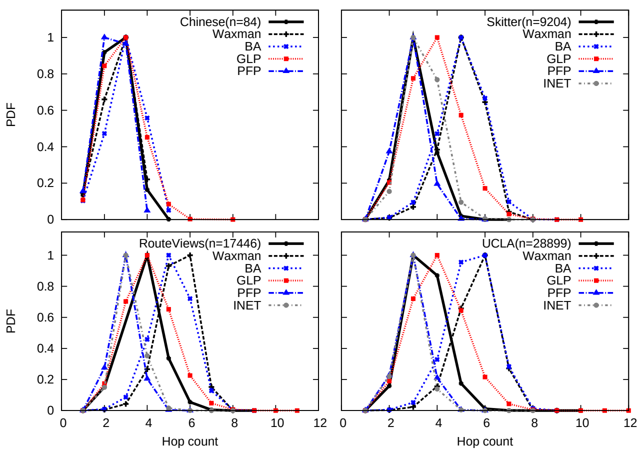

Fig. 5: Comparison of shortest path distributions (number of hops).

For medium degree nodes, clustering coefficients display rather high variability which increases with the size of the observed topologies. This behavior is 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 underestimating it for high-degree nodes. Current topology generators do not adequately model local node connectivity.

4) Rich-club connectivity: Rich-club connectivity gives information about how well-connected nodes of high degree are among themselves. Figure 4 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. Waxman and BA perform poorly for this metric. 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.

5) Shortest path distributions: Figure 5 displays the distributions of shortest path length. Apart from BA, 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 capture the path length distribution for particular topologies well: PFP matches Skitter's well and GLP is close for Routeviews. Inet and PFP both focus on high connectivity in the core of the network, hence they both match UCLA better than 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.

6) Weighted Spectral Distribution: The Weighted Spectral Distribution (WSD) was initially introduced in [23] and further expanded upon in [10], [24]. It is based on the eigenvalues (i.e. spectrum) of the normalized Laplacian matrix of a graph. As shown in [10] the difference between the WSD's of two graphs forms a distance metric, i.e. two graphs may have the same WSD only if they are equal and also it can be used to determine which of two (or more) graphs is closer to a target graph. The WSD is composed of a curve (a weighted distribution) parametrized by an integer N. The curve is essentially the power in each cluster of the graph that contributes to the probability of taking a *random N-cycle walk* on a graph. For example, a random 4-cycle walk ($N = 4$) is a random walk starting and ending at the same node having passed 2 nodes in-between ($a \to b \to c \to a$). The probability of taking *any* such walk on a graph is simply the sum of the WSD curve. The contribution of each cluster in the graph to this sum is the WSD and is unique to each graph. Thus in a very useful sense the WSD represents the structure of a graph [10].

Figure 6 displays the WSD of the Skitter data set and the closest WSD that each topology generator is able to obtain. First note that no topology generator achieves the same WSD as Skitter. This indicates that there is more structure in the observed graph than can be accounted for by any of these models. In addition note that PFP obtains the best fit followed closely by the BA and GLP generators. The Waxman generator

[10] The WSD is *(i) self-replicating*, i.e. The WSD can be used to estimate the (unknown) parameters of a graph of given type (for example BA); *(ii) monotonic*; as the estimated parameters deviate from the true parameters the WSD distance increases and *(iii) unique*; the WSD's of (for example) an BA type graph and GLP type graph cannot agree.

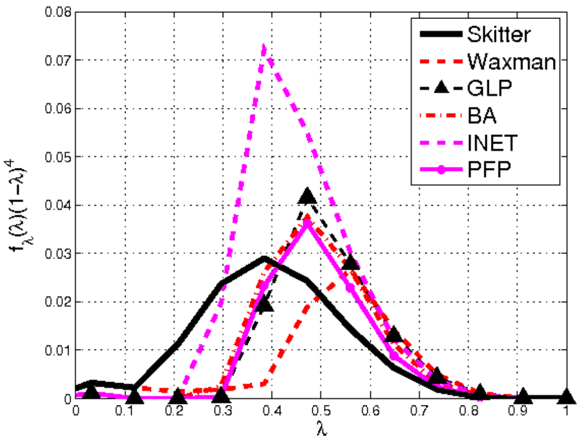

Fig. 6: Best fit WSDs for topology generators relative to target Skitter data set.

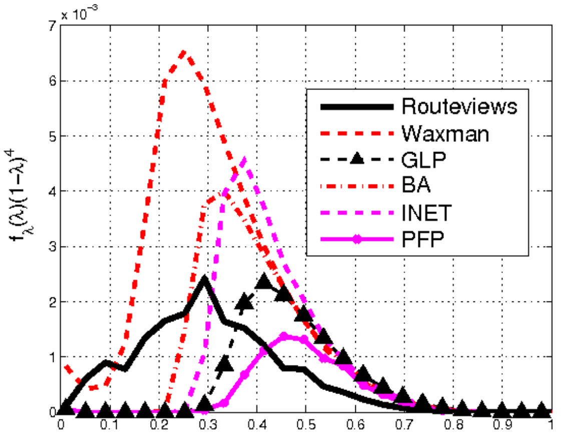

Fig. 7: Best fit WSDs for topology generators relative to target Routeviews data set.

obtains the worst fit due its random graph model that is a poor fit for the Internet. The INET model is interesting in that it achieves its maximum at the right point ($\lambda = 0.4$), but the power at this point is too high. This is an artifact of the simple way in which the core is constructed in INET, producing many more 4-cycles than seen in the observed data set.

Figure 7 displays the Routeviews data set with the best WSD fit obtainable from each of the topology generators. Again none of the topology generators obtains a good fit. The Waxman generator again performs worst. Based on the sum squared error fit PFP performs best followed by GLP, BA and INET although the differences are small between them.

C. 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 led 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. We suggest that the topology generators should focus more on metrics such as clustering and WSD for tuning and optimizing topology generators [23].

V. Multiple vantage points

The previous section studied in detail *how well* topology generators capture the properties of different observed AS topologies. In this section, we will study *why* topology generators capture different properties of observed AS topologies with varying degrees of success. To that end we examine the impact on the metrics of the number of vantage points from which BGP data is collected. For our analysis we collected BGP data from over 40 RouteViews peering points, for a period of 6 months from May 2007. This time period was chosen to be the same as that used to build the UCLA dataset.

Table II shows the values of the topological metrics the same way as in Table I, for AS topologies obtained from different numbers of observation points. When comparing the AS topologies using 1 and 10 observation points, we see a significant increase in the number of nodes and links. BGP observation points typically see a limited fraction of the AS links, and even a subset of the nodes as the first peer on Table I. Hence, one might also expect a significant difference in the other metrics, and indeed, the maximum node degree almost triples and the number of fully-meshed nodes almost doubles. As a consequence, the size of the core increases, indicated by the maximum coreness value. In turn, the number of shortest paths crossing the core also increases as indicated by the maximum betweenness. On the other hand, going from 1 to 10 observation points slightly decreases the value of the clustering coefficient. This is because those observation points lie in the core of the network and represent the path diversity in the core. Having different observation points in the edge of the network would show different results, however such data is not available today. With 25 or more observation points the links on the edge of the network are also discovered, contributing to the increase of the value of the clustering coefficient. This behavior is confirmed by a slight decrease of the value of the maximum betweenness from 10 to 25 observation points.

Preferential attachment models originate in the belief that small ASs tend to connect to large upstream ASs, leading to a disassortative network. Although the value of the assortativity coefficient is negative for the AS topology, it is not affected by an increase in the number of observation points. The links added by increasing the number of observation points are neutral for the assortativity of the AS topology. One implication is that the links that can be discovered by using more observation points do not preferentially interconnect ASs of any particular degree.

Our conjecture is that the observation points added from RouteViews do not preferentially miss peer-peer relationships

TABLE II: Comparison of AS topology datasets from multiple peering points.

Topology	Nodes	Links	Avg. deg.	Max. degree	Top clique size	Max. betweenness	Max. coreness	Assort. coef.	Clust. coef.	Max. closeness
1 peer	17,952	34,617	3.86	980	4	35,069,182	9	-0.18	0.008	<0.01
10 peers	27,838	64,717	4.65	2,731	7	52,862,315	20	-0.18	0.007	<0.01
25 peers	27,885	67,659	4.85	2,808	7	49,798,002	25	-0.19	0.01	<0.01
All peers	27,924	70,064	5.02	3,371	7	70,142,726	30	-0.18	0.01	<0.01

because of the current poor visibility of peer-peer relationships from core ASs. RouteViews sees the Internet mostly from its core, not the edge. Other sources of measurements (e.g., traceroutes) or BGP observations from different types of ASs may reveal a different Internet structure [25], especially at the edge where many peer-peer relationships might be hidden. Some note of caution is necessary though. The process of discovering new AS edges by adding observation points does not have to reflect how many edges are actually not seen by BGP [25].

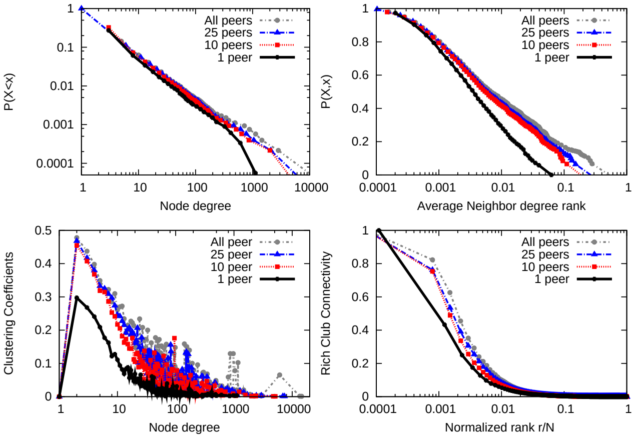

Fig. 8: Comparison of effects of the number of peering points.

We now turn in more detail to the effect of the number of peering points on four topological metrics (see Figure 8). The addition of observation points mostly affects node degree distribution for high degree nodes. As we increase the number of observation points, on average the neighbors of a node will have a higher degree. However, this does not hold for nodes whose neighbors already have high degrees (left part of Figure 8). Those nodes correspond to stub networks connected to very well interconnected upstream providers. For the clustering coefficient, when moving from one to several observation points, the difference is striking. For all node degrees, the clustering coefficient significantly increases. On the other hand, when moving from a few peerings to many, the difference appears mostly for high degree nodes. This illustrates the better observability of links in the core compared to the edge of the network. Rich-club connectivity confirms the previous observations in that adding a few observation points is enough to discover the core links.

In this section we have illustrated the importance of relying on a sufficiently large number of observation points in order to capture a wider set of properties of the AS topology. Using only a few observation points has led researchers to simplify the complexity of the interconnection structure between ASs. Taking observations of the AS topology at face value is dangerous [25], as researchers are still trying to understand the actual properties of the AS topology. For example, the types and numbers of AS edges that are missed remain open issues [25]–[27]. From our study, it is questionable whether it is actually possible to argue about the "true" properties of the AS topology. Proposing new AS topology models thus faces the problem of availability of representative datasets. Our results show that researchers must use rich datasets for a proper understanding of the Internet AS topology. How much better than today's publicly available data is necessary to better understand the AS topology is debatable.

VI. Related Work

Zegura *et al.* [28] analyse topologies of 100 nodes generated using pure-random, Waxman [11], exponential and several locality based models of topology such as Transit-Stub. They use metrics, such as average node degree, network diameter, number of paths between nodes. They find that pure random topologies represent properties such as locality very poorly and so we exclude them from our comparison. They suggest that the Transit-Stub method should be used due both to its efficiency and the realistic average node degree its topologies achieve.

Faloutsos *et al.* [8] state that three specific properties of the Internet AS topology are well described by power laws: rank exponent, out-degree exponent and eigen-exponent (graph eigenvalues). This work parallelled development of many models incorporating power laws based on preferential attachment, e.g., the Barabási and Albert [15] model.

Bu and Towsley [5] compare the effectiveness of several topology generators at creating power law topologies that model the AS topology. They show that existing topology generators capture well the power law exponent, but fail to capture clustering properties and path length. They propose a new topology generator, GLP, based on preferential attachment.

Tangmunarunkit *et al.* [29] provide the first comparison of degree-based models and structural models. They compare three categories of model generators: Waxman, Tiers [30] and the Transit-Stub structural model, against the simplest degree based generator, the power-law random graph (PLRG) [31]. They find that the PLRG matches these metrics better than random or structural models. They conclude that a stricter hierarchy is present in the measured networks than in degree-based generators. They also conclude that the simplest form of degree-based model performs better than random or structural models.

VII. Conclusions

In this paper we evaluated some existing AS topology generation models, by comparing them with several observed AS topologies. For this evaluation, we relied on a wide set of topological measures, including the graph spectrum, to carry our comparison as objectively as possible. Our analysis revealed that increasing the number of observation points causes deviation from strict degree power-law scaling. Existing topology generation models over emphasize the preferential attachment mechanism and the resulting node degree distribution. Strict power-law scaling appears to be an artifact of incomplete datasets, rather than a fundamental property of the AS topology.

In addition to clustering and centrality properties, we observe that the highly meshed core of the Internet AS topology must be included to generate representative synthetic topologies. The successive improvements in topology generation models seem to result from improvements in the available datasets. Knowing that incomplete datasets were the cause for simplistic topology generation models, we expect that the new generation of topology models will take into account the insights gained in this paper.

The main insights of this paper concern the importance of observations of the AS topology on the current assumptions about its topological properties. Improving the representativeness of the available data is crucial to properly understand the topological properties of the Internet. As we show in this paper, it is most likely because of local structural properties that additional data is necessary. Our insights indicate that additional measurements should come from the edge of the network to improve our understanding of the properties of the AS topology.

References

[1] B. Halabi, *Internet Routing Architectures.* Cisco Press, 1997.

[2] Y. Rekhter, T. Li, and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” IETF, RFC 4271, Jan. 2006.

[3] B. Huffaker, D. Plummer, D. Moore, , and k Claffy, “Topology discovery by active probing,” in *SAINT-W’02.* IEEE Computer Society, 2002, p. 90.

[4] H. Haddadi, G. Iannaccone, A. Moore, R. Mortier, and M. Rio, “Network topologies: Inference, modelling and generation,” in *IEEE Communications Surveys and Tutorials*, vol. 10, no. 2, 2008.

[5] T. Bu and D. Towsley, “On distinguishing between Internet power law topology generators,” in *Proceedings of IEEE Infocom 2002*, New York, NY, Jun. 2002.

[6] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, “Systematic topology analysis and generation using degree correlations,” in *Proceedings of ACM SIGCOMM 2006*, Pisa, Italy, 2006, pp. 135–146.

[7] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz, “Characterizing the Internet hierarchy from multiple vantage points,” in *Proceedings of IEEE Infocom 2002.* New York, NY: IEEE, Jun 2002.

[8] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the Internet topology,” in *Proceedings of ACM SIGCOMM 1999*, Cambridge, Massachusetts, United States, 1999, pp. 251–262.

[9] 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,” University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-725, Jul. 2008.

[10] D. Fay, H. Haddadi, S. Uhlig, A. W. Moore, R. Mortier, and A. Jamakovic, “Weighted spectral distribution,” University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-729, Sep. 2008. [Online]. Available: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-729.pdf

[11] B. M. Waxman, “Routing of multipoint connections,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 6, no. 9, pp. 1617–1622, Dec. 1988.

[12] P. Erdos and A. Renyi, “On random graphs,” in *Mathematical Institute Hungarian Academy, 196.* London: Academic Press, 1985.

[13] A. Medina, A. Lakhina, I. Matta, and J. Byers, “BRITE: an approach to universal topology generation,” in *IEEE MASCOTS*, Cincinnati, OH, USA, Aug. 2001, pp. 346–353.

[14] R. Albert and A.-L. Barabasi, “Topology of evolving networks: local events and universality,” *Physical Review Letters*, vol. 85, p. 5234, (2000). [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0005085

[15] A. L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[16] J. Winick and S. Jamin, “Inet-3.0: Internet topology generator,” University of Michigan, Tech. Rep. CSE-TR-456-02, 2002.

[17] S. Zhou and R. J. Mondragn, “Accurately modeling the internet topology,” *Phys. Rev. E*, vol. 70, 2004.

[18] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs, “Locating Internet routing instabilities,” in *Proceedings of ACM SIGCOMM 2004*, Portland, OR, 2004, pp. 205–218.

[19] Z. M. Mao, J. Rexford, J. Wang, and R. H. Katz, “Towards an accurate AS-level traceroute tool,” in *Proceedings of ACM SIGCOMM 2003*, Karlsruhe, Germany, 2003, pp. 365–378.

[20] S. Zhou, G.-Q. Zhang, and G.-Q. Zhang, “Chinese Internet AS-level topology,” *IET Communications*, vol. 1, no. 2, pp. 209–214, April 2007.

[21] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, k c claffy, and A. Vahdat, “The Internet AS-level topology: three data sources and one definitive metric,” *SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 17–26, 2006.

[22] R. Oliveira, B. Zhang, and L. Zhang, “Observing the Evolution of Internet AS Topology,” in *Proceedings of ACM SIGCOMM 2007*, Kyoto, Japan, Aug. 2007.

[23] H. Haddadi, D. Fay, S. Uhlig, A. Moore, R. Mortier, A. Jamakovic, and M. Rio, “Tuning topology generators using spectral distributions,” in *LNCS, SIPEW*, vol. 5119. Darmstadt, Germany: Springer, 2008.

[24] D. Fay, H. Haddadi, A. Thomason, A. W. Moore, R. Mortier, A. Jamakovic, S. Uhlig, and M. Rio, “Weighted Spectral Distribution for Internet Topology Analysis: Theory and Applications,” *IEEE/ACM Transactions on Networking (TON)*, To Appear.

[25] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, “In search of the elusive ground truth: The Internet’s AS-level connectivity structure,” in *ACM SIGMETRICS*, Annapolis, USA, Jun. 2008.

[26] H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, “Towards capturing representative AS-level Internet topologies,” *Computer Networks*, vol. 44, no. 6, pp. 737–755, 2004.

[27] Y. He, G. Siganos, M. Faloutsos, and S. Krishnamurthy, “A systematic framework for unearthing the missing links: Measurements and impact,” in *Proceedings of Third Symposium on Networked Systems Design and Implementation (NSDI) 2007*, 2007, pp. 187– 200.

[28] E. W. Zegura, K. L. Calvert, and M. J. Donahoo, “A quantitative comparison of graph-based models for internet topology,” *IEEE/ACM Transactions on Networking (TON)*, vol. 5, no. 6, pp. 770–783, 1997.

[29] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, “Network topology generators: degree-based vs. structural,” in *Proceedings of ACM SIGCOMM 2002*, Pittsburgh, PA, 2002, pp. 147–159.

[30] M. B. Doar, “A better model for generating test networks,” in *IEEE GLOBECOM’96*, London, UK, Nov. 1996.

[31] W. Aiello, F. Chung, and L. Lu, “A random graph model for massive graphs,” in *STOC’00*, Portland, OR, May 2000, pp. 171–180.