

# Comparative Analysis of the Internet AS-Level Topologies Extracted from Different Data Sources

Priya Mahadevan  
UCSD/CAIDA  
priya@caida.org

Dmitri Krioukov  
CAIDA  
dima@caida.org

Bradley Huffaker  
CAIDA  
brad@caida.org

Xenofontas Dimitropoulos  
Georgia Tech  
fontas@ece.gatech.edu

kc claffy  
CAIDA  
kc@caida.org

Amin Vahdat  
UCSD  
vahdat@cs.ucsd.edu

## Abstract

We calculate vital statistics of Internet AS-level topologies extracted from the four data sources used by the research community: traceroute measurements, BGP tables, BGP updates, and WHOIS database. We find that topologies derived from BGP tables and BGP updates are virtually identical, but differ from the traceroute topology in terms of the vital statistics that we present in this paper. The WHOIS topology is substantially different from the three other topology graphs. We introduce the concept of *dimensionality of graph randomness* to evaluate the accuracy of topology generators. We show that power-law random graph generators cannot capture the full spectrum of critical topological characteristics of traceroute- and BGP-derived graphs. However, the random graph model reproducing the *joint degree distribution* of the respective graphs (traceroute, BGP, WHOIS) better approximates the actual topology for traceroute data than for the other data sources. These observations have direct implications for the development of new and credibility of existing network topology generators. Finally, we release to the community both the input topologies and the output of our calculations, which will allow researchers to make a more informed selection of topology data for their specific needs.

## 1 Introduction

Internet topology analysis and modeling has attracted substantial attention recently [26, 52, 6, 7, 56, 18, 53, 17, 39, 8, 34, 27, 58]<sup>1</sup>, which is not surprising since the Internet’s topological properties and their evolution are cornerstones of many practical and theoretical network research agendas. Our own leading motivation is the need to construct accurate network emulation environments to support development, reliable testing, and performance evaluation of new applications, protocols, and routing architectures [15]. Realistic network topologies, and tools to generate them, are essential to this goal.

Given a synthetic topology produced by a topology generator, we seek to compare it with a measured topology; or, if there are multiple sources of topological data (the case we have in this paper), we seek to compare several *measured* topologies. The adjacency matrix (cf. Section 3.2.4) of the graph representing a given network contains full information on its topology. Unfortunately, it is impossible to compare topologies based on their adjacency matrices, since two similar  $n \times n$ -matrices might represent two graphs with completely different topological properties. The most basic set of graph properties is the set of its *connectivity* characteristics.

As a big surprise to many came a revelation in [26] that the Internet *connectivity distribution* (or node degree distribution, cf. Section 3.1.2) follows a power law. Later work [53] demonstrated that topology generators blindly reproducing the observed power-law node degree distribution deliver more realistic topologies than the generators trying to capture the perceived hierarchical properties of the Internet. The authors of [53] distilled a set of characteristics to differentiate between measured and synthetic topologies. These characteristics were: (1) *expansion*, closely related to the graph distance distribution (cf. Section 3.2.1), (2) *resilience*, related to the minimum balanced cut of a graph, which we can estimate by its largest eigenvalue (cf. Section 3.2.4), and (3) *distortion*, the minimum stretch of a tree spanner, which we do not discuss in this paper. The work in [53] also analyses the Internet hierarchy by calculating the *link value* distribution, where link value is its *betweenness* (cf. Section 3.2.3).

<sup>1</sup>We intentionally avoid citing statistical physics literature, where the number of publications dedicated to the subject has exploded. For introduction and references see [23, 15, 36].

In recent interesting work [39], the authors perform an even finer analysis of the network topology by: (1) evaluating the network’s performance using betweenness-related characteristics, and (2) calculating graph *likelihood*, which turns out to be closely related to the network assortativity coefficient—the summary statistics of the *joint degree distribution*. The joint degree distribution provides a much finer description of a graph’s connectivity than its degree distribution (cf. Section 3.1.3). It is important to note that the authors of [39] discuss the *router-level* topology.

Internet topologies can be described at various levels of granularity, most notably at the *AS* (Autonomous System) and *router* levels. In this paper we focus on AS-level topologies, due to their relevance to interdomain routing, which frames a set of engineering and economic concerns that directly affect the resilience and robustness of the network.

There are at least four different sources of Internet AS-level topology data, measured using four different methodologies and yielding substantially different characteristics. Unfortunately many researchers rely only on one data source, sometimes outdated or incomplete. Another common technique is to mix disparate data sources into one topology, but there has been no attempt to provide a detailed analytical comparison of the vital statistics of the different topology data sources taken separately.

In this paper we fill this gap by calculating the most basic and commonly used topological properties of Internet AS-level graphs extracted from four data sources: (1) traceroute measurements; (2) BGP tables; (3) BGP updates; and (4) WHOIS. This work has four key contributions to the field of topology research:

1. A comprehensive set of calculated metrics describing structural properties of Internet graphs.
2. An exhaustive comparison of AS-level graphs obtained from different data sources in terms of these metrics, including an exposition of interdependencies among calculated statistics for the four classes of data.
3. Introduction of the concept of *dimensionality of graph randomness* to help evaluate how well random graph generators based on different principles approximate the four observed topologies. We believe these observations will impact the development of new network topology generators.
4. Release of data and results to the community [11]: a) the input graphs representing the topologies extracted from the four raw data sources; b) the data plots (many not included in the paper) of the results of our calculations; c) the data files associated with the plots, useful for researchers seeking other summary statistics; and d) the scripts and programs we developed for our calculations, useful for researchers looking for a tool to validate the veracity of a given topology or generator.

We cannot yet say with conviction whether a specific measured topology is closest to the real AS-level Internet topology, but our comparison should arm researchers with better insight into specifics of each topology and allow more informed selection of topology data for their needs.

We organize this paper as follows. Section 2 describes the four data sources and how we prepare the data to obtain the four input graphs. Section 3 lists and defines the set of topological characteristics we calculate, explaining what they measure and why they are important. Section 4 compares vital statistics for the four topology graphs. We conclude in Section 5 with the summary of our findings.

## 2 Data sources

We calculate topology characteristics of the Internet AS-level graphs derived from the following four data sources: traceroute measurements, BGP tables, BGP updates, and WHOIS database.

**Traceroute** [54] is a tool that captures a sequence of IP hops along the forward path from the source to a given destination by sending packets (UDP or ICMP probes) to the destination. Among the four data sources, traceroute measurements are most likely to reflect the topology of actual Internet traffic flows, i.e., the data plane.

Under the Macroscopic Topology Project [13], CAIDA has developed significant infrastructure for continuous traceroute-based Internet topology measurements collected by *skitter* [35]. CAIDA also provides AS-level topology graphs derived from the skitter data aggregated on a daily basis [12]. We filter out AS-sets [49], multi-origin ASs [57], private ASs [30], and indirect links [12] from the data provided at [12]. We then merge the 31 daily AS graphs for the month of March 2004.

**BGP** (Border Gateway Protocol) [49] is the protocol used for routing among ASs in the Internet. BGP data provides the topology view seen by the routing system (i.e., control plane) of the Internet. Therefore, topology extracted from BGP data is less likely to accurately reflect the topology seen by the data plane. Detailed characterization of incongruities between traceroute- and BGP-derived topologies is a subject of on-going research [31, 43, 42].

RouteViews [50] collects and archives both **BGP tables** (static snapshots of the BGP routing tables) and **BGP updates** (dynamic BGP data in the form of actual control packets, i.e., updates and withdrawals, used for routing table construction). We extract AS links from RouteViews tables using the **StraightenRV** script from CAIDA’s **rv2atoms** package [14]. We filter out AS-sets and private ASs. We merge daily AS graphs for the same month of March 2004. For the BGP-update-derived topology, we extract AS links from BGP update messages accumulated over the same period (March 2004) and filter the resulting data the same way.

**WHOIS** [48] is a collection of databases containing a wide range of information useful to network operators. Manually maintained with no requirements for updating the registered information, these databases are even less accurate than BGP in depicting the data plane topology. The only WHOIS database containing relatively reliable current topological information is RIPE’s [51, 17], which covers primarily European Internet infrastructure.

We obtain the RIPE WHOIS database dump for April 07, 2004. The database records of interest to us look like

```
aut-num: ASx
import: from ASy
export: to ASz
```

which indicate links **ASx-ASy** and **ASx-ASz**. After constructing the AS-level graph from this data, we exclude ASs that do not appear in the **aut-num** lines. Such ASs are external to the database and we cannot correctly estimate their topological properties (e.g., node degree). Finally, we filter out private ASs.

All four graphs constructed as described are available for download from [11].

### 3 Topology characteristics

We conducted an exhaustive review of literature in Internet topology analysis and compiled a set of metrics that are most basic and commonly used. While we do not claim this set to be complete, we believe it captures most important properties of Internet topology. It frames our comparison of structural properties of the Internet topologies derived from the four data sources described in Section 2.

We split the metrics in this set into two categories: metrics describing local connectivity in a graph and characteristics of the global structure of the topology.

#### 3.1 Local connectivity

##### 3.1.1 Average degree

The two most basic graph properties are the **number of nodes**  $n$  (also referred as *graph size*) and the **number of links**  $m$ . They define the coarsest connectivity characteristic, which is the **average node degree**  $\bar{k} = 2m/n$ .

Given  $n$  and  $m$  (and consequently  $\bar{k}$ ), one can construct the class of maximally random graphs having the required average degree. We call such graphs **K-random**. These are classical random graphs  $G_{n,m}$  [25], which are substantially different from realistic Internet topologies [23].

The average node degree has limited utility since graphs with the same average node degree can have vastly different topological structures.

##### 3.1.2 Degree distribution

A more informative characteristic of graph connectivity is the **node degree distribution**, which provides information on the number of nodes in the graph having a certain degree. If  $n(k)$  is the number of nodes of degree  $k$  ( $k$ -degree nodes) in a graph, then the node degree distribution is the probability that a randomly selected node has degree equal to  $k$ :  $P(k) = n(k)/n$ .

The degree distribution contains more information about connectivity in a given graph than the average degree, since given a specific form of  $P(k)$  we can always restore the average degree by  $\bar{k} = \sum_{k=1}^{k_{max}} kP(k)$ , where  $k_{max}$  is the **maximum node degree** in the graph.

Note that the K-random graphs have a specific form of the degree distribution—the binomial distribution, which we can closely approximate by the Poisson distribution [23]:  $P_K(k) = e^{-\bar{k}} \bar{k}^k / k!$ .

As was first observed in [26], the node degree distribution in the Internet is close to the power law,  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is a positive **exponent** of the power-law node degree distribution. If the node degree distribution in a graph follows the power law exactly, then the distribution has a natural cut-off at the **power-law maximum degree** [23]:  $k_{max}^{PL} = n^{1/(\gamma-1)}$ .

Given  $P(k)$ , one can construct the class of maximally random graphs having the required degree distribution by following a recipe introduced in [44, 45] and further formalized in [1]. We call such graphs **PK-random** or *power-law random graphs* (PLRG). As demonstrated in [53], PLRG-based topology generators produce more veracious results than the older Internet topology generators designed to reproduce the perceived hierarchical structure of the Internet.

However, as we will see in Section 4, the topology generation strategy based on reproducing only the degree distribution fails to capture a wide spectrum of other important topological properties. Indeed, the degree distribution is a microscopic connectivity characteristic in the following sense: it tells us much more than the average node degree, but it does not tell us anything about how these nodes are interconnected.

### 3.1.3 Joint degree distribution

The natural extension of the degree distribution is the **joint degree distribution**  $P(k_1, k_2)$ , which tells us how nodes interconnect: the values of  $P(k_1, k_2)$  give the probability that a randomly selected edge has degrees of adjacent nodes equal to  $k_1$  and  $k_2$ . Note that this probability is different from the conditional probability  $P(k_2|k_1)$  that a given node of degree  $k_1$  is connected to a node of degree  $k_2$  [23].

The joint degree distribution contains more information about the connectivity in a graph than the average degree, since given a specific form of  $P(k_1, k_2)$  we can always restore both the degree distribution and average degree by expressions that one can find in [23].

Note that the PK-random graphs have a specific form of the joint degree distribution [23]. If we denote by  $\tilde{P}(k)$  the probability that one of two nodes adjacent to a randomly selected edge is of degree  $k$ ,  $\tilde{P}(k) = (k/\bar{k})P(k)$ , then the joint degree distribution in PK-random graphs is  $P_{PK}(k_1, k_2) = \tilde{P}(k_1)\tilde{P}(k_2)$ , meaning that there are no correlations between degrees of adjacent nodes. This is why PK-random graphs are also called *uncorrelated* [23]. According to the construction [25], K-random graphs are also uncorrelated, with  $P(k_1, k_2)$  given by the same expression above, except that  $P(k)$  is the Poisson distribution  $P_K(k)$  from Section 3.1.2.

Given  $P(k_1, k_2)$ , one can construct the class of the maximally random graph having the required *joint degree distribution* by following a recipe introduced in [46]. We call such graphs **PKK-random**.

The joint degree distribution, being a function of two sparse discrete arguments, is hard to depict, although we try to do so in Section 4. Easier to plot is the summary statistic of the joint degree distribution called **the average neighbor connectivity**  $k_{nn}(k)$ . As the name suggests, the average neighbor connectivity is the average degree of the nearest neighbors of nodes of degree  $k$ . We can calculate it by  $k_{nn}(k) = \sum_{k'=1}^{k_{max}} k'P(k'|k)$ . Note that in uncorrelated networks, it is constant  $k_{nn}^{PK}(k) = \langle k^2 \rangle / \bar{k}$  [23], where by  $\langle k^2 \rangle$  we denote the second raw moment of the degree distribution. In the K-random graph case, we have  $k_{nn}^K(k) = \bar{k} + 1$ . For better graph comparison, it is convenient to normalize  $k_{nn}(k)$  by dividing it by its maximal possible value  $n - 1$  (cf. Section 4).

We can further summarize the joint degree distribution statistics by means of the **assortativity coefficient** introduced in [46] and redefined as *likelihood* in a recent study [39]. The value of  $r = 0$  corresponds to uncorrelated PK-random graphs. Graphs with  $r > 0$  are **assortative**. They have an excess of links interconnecting high-degree nodes compared to PK-random graphs. Inspired by the design of the skitter poster [16], where higher degree nodes are placed closer to the center of a circular diagram, we call links interconnecting nodes of similar degrees **tangential**, while links connecting higher-degree nodes to lower-degree nodes are **radial**. In this terminology, assortative graphs have an excess of tangential links in the core, the connectivity inside the core is richer, and as a consequence, assortative graphs are harder to break (e.g., to decompose into similarly sized disconnected components). Classic examples are social networks [46].

As we will see in Section 4, all four of our data sources imply that the Internet is an example of **disassortative** graphs with  $r < 0$ . Such graphs have more radial links than in the PK-random case. In other words, the connectivity inside the core is sparser and as a consequence such graphs are easier to break.

### 3.1.4 Clustering

While the average neighbor degree suggests whether an AS is connected to high-degree or low-degree ASs, it fails to tell us how these neighbor ASs are interconnected. Clustering captures how close the neighbors of an AS are to forming a clique. In other words, this metric provides even more detailed information on the local structure of a graph than the joint degree distribution. Clustering also measures local robustness in the graph—the higher the local clustering of an AS node, the more interconnected are its neighbors, increasing the existence of alternate paths locally around the AS.

More formally, if we denote by  $\langle m_{nn}(k) \rangle$  the average number of links between the nearest neighbors of  $k$ -degree nodes, then **local clustering** is the ratio of this number to the maximum possible number of such links:

$C(k) = \langle m_{nn}(k) \rangle / (k(k-1)/2)$ . Note that this definition also tells us the average number of cycles of length 3 (3-cycles) involving  $k$ -degree nodes. The form of function  $C(k)$  provides information on the hierarchical structure of the graph. For example, smaller values of  $C(k)$  for lower-degree nodes than for high-degree nodes tell us that clusters of small ASs tend to interconnect via relatively less clustered high-degree hubs.

There are two summary statistics associated with local clustering: the **mean local clustering**  $\bar{C}$ , which is just the average value of  $C(k)$ , and the **clustering coefficient**  $C$  which is a normalized measure (weighted by the number of connected node triplets) of the number of loops of length 3 in a graph [4].

As shown in [22], the PKK-random graphs have a specific form of local clustering, Eq. (8) in [22], which we denote by  $C_{PKK}(k)$ . It defines both the PKK-random mean local clustering  $\bar{C}_{PKK}$  and clustering coefficient  $C_{PKK}$  (Eqs. (9) and (10) in [22]). When compared with clustering observed in the real Internet, clustering in the PKK-random graphs allows us to estimate how far in terms of neighbor (1-hop) connectivity the measured topologies are from the PKK-random topologies. In other words, we have a measure of accuracy of approximation of real topologies by PKK-random ones.

Note that uncorrelated PK-random graphs have not only constant average neighbor connectivity but also constant clustering [22]:  $C_{PK}(k) = \bar{C}_{PK} = C_{PK} = (\langle k^2 \rangle - \bar{k}^2) / (n\bar{k}^3)$ , which clearly differs from the PKK-random case. In the K-random graph case, we have [23]:  $C_K(k) = \bar{C}_K = C_K = \bar{k}/n$ .

### 3.1.5 Rich club connectivity

As of this writing, the Internet AS-level topology model most accurately reproducing the largest set of important topology characteristics is the Positive-Feedback Preference (PFP) model introduced in [58]. Interestingly, this model is successful in reproducing the measured AS-level graph by trying to capture only the following three characteristics: the exact form of the node degree distribution; the maximum node degree; and the rich club connectivity, which the authors of [58] define as follows.

Rank nodes by their degrees in the non-increasing order. Let  $\rho = 1 \dots n$  be the first  $\rho$  nodes ordered by their rank. **Rich club connectivity**  $\phi(\rho/n)$  is the ratio of the number of links in the graph induced by these  $\rho$  nodes to the maximum possible number of links  $\rho(\rho-1)$ . Defined this way,  $\phi(\rho/n)$  is a measure of how close the subgraph induced by the  $\rho$  largest degree nodes (the  $\rho$ -induced subgraph) is to a clique.

We also calculate the **top clique size**  $n_{clique}$ : the maximum number of highest rank nodes still forming a clique. This number tells us how well-connected the top degree nodes are.

It is still unclear if the rich club connectivity property reveals yet another “degree of freedom” of connectivity characteristic of a graph. It is certainly not redundant with the assortativity coefficient [58]. The success of the PFP model in capturing so many characteristics of the real topology has yet to be explained. For now we use it as a convenient measure of cliquishness of  $\rho$ -induced subgraphs. In particular, when  $\rho$  is small, the rich club metric describes connectivity in the core, although a better measure of core connectivity is coreness.

### 3.1.6 Coreness

There are two definitions of coreness. In graph-theoretic literature [3], the  $k$ -core of a graph is the subgraph obtained from the original graph after removal of all nodes of degree less than or equal to  $k$ . A more informative definition of the  $k$ -core [27] is the subgraph obtained from the original by *iterative* removal of all nodes of degree less than or equal to  $k$ .

The **node coreness**  $\kappa_i$  of node  $i$  is then the maximum  $k$  such that  $i$  is still present in the  $k$ -core but removed in the  $(k+1)$ -core. All 1-degree nodes have  $\kappa = 0$ . The **graph coreness**  $\kappa_{max}$  is the maximum node coreness in a graph. By definition, the  $\kappa_{max}$ -core is not empty, but  $(\kappa_{max} + 1)$ -core is.

We further define the **graph core** to be its  $\kappa_{max}$ -core, and the **graph fringe** to be the set of nodes with minimum coreness  $\kappa_{min}$ . If  $\kappa_{min} = \kappa_{max}$ , then the graph’s fringe and core coincide with the whole graph.

By construction, the graph coreness characterizes the richness of connectivity in the graph. Coreness of a tree is 0. Coreness of a  $k$ -regular graph [29] is equal to coreness of all of its nodes (all having degree  $k$ ), which is equal to  $k-1$ .

The node coreness tells us how “deep in the core” the node is. It is interesting to relate this measure to the most basic connectivity characteristic of a node: its degree. We can do this comparison by considering the average node coreness as a function of node degree  $\kappa(k)$ . Note that node coreness is a much more sophisticated measure of node connectivity than its degree. Indeed, the node degree can be high, but if its coreness is small, then the node is not well connected, we can easily disconnect it by removing its poorly connected neighborhood. For example, the high-degree hub of a star has coreness of 0.

At the same time, node coreness is not a measure of centrality of the node. Indeed, a low-degree node interconnecting a few high-degree hubs has a low value of coreness, but intuitively it is in the “center of the graph.” To measure centrality, we need some other characteristics that are functions of the global structure of the graph.

## 3.2 Global structure

### 3.2.1 Distance

On the other end of our increasingly macroscopic characterization of graph connectivity (i.e.,  $\bar{k}$ ,  $P(k)$ ,  $P(k_1, k_2)$ ) lies the shortest path length distribution, which we call simply the **distance distribution**  $d(x)$ , which is the probability for a random pair of nodes to be at distance of  $x$  hops from each other. Interestingly, the distance distribution does not tell us anything about connectivity of nodes, but the node connectivity distribution  $P(k)$  of PK-random graphs defines their distance distribution [24]. Unfortunately, there are no results of this nature for more complicated cases, such as the PKK-random case.

The two basic summary statistics associated with the distance distribution are the **average distance**  $\bar{d}$  in a graph and the **standard deviation**  $\sigma$  of its distance distribution. We can call the latter the *distance distribution width* since as Section 4 will show, the distance distribution in the Internet (and many other networks) has a characteristic Gaussian-like form.

We can use distance as a measure of node centrality. Indeed, nodes with smaller average distances are closer to the graph “center.” Similar to the case with coreness, we can relate this measure to node degree by considering the average distance from  $k$ -degree nodes  $d(k)$ .

Beyond providing the most basic measure of node centrality, the distance distribution is critical for many applications, the most prominent being routing. The distance-based locality-sensitive approach [47] lies at the root of virtually all modern routing algorithms, and it is not hard to see that performance parameters of these algorithms depend strongly on the distance distribution of a graph [37]. In particular, short average distance and small distance distribution width render routing especially difficult.

### 3.2.2 Eccentricity

Eccentricity is an extremal form of distance [29]. If  $d_{ij}$  is the distance between nodes  $i$  and  $j$ , then **eccentricity**  $\varepsilon_i$  of node  $i$  is the maximum distance from  $i$ :  $\varepsilon_i = \max_j d_{ij}$ . The maximum distance  $D$  in a graph, its **diameter**, is also the maximum eccentricity,  $D = \max_i \varepsilon_i = \max_i \max_j d_{ij}$ , and the minimum eccentricity  $R$  is called the graph **radius**,  $R = \min_i \varepsilon_i = \min_i \max_j d_{ij}$ .

Eccentricity is another measure of centrality of the node, except that it considers the maximum distance instead of average distance. It is important for practical applications, since we need to be able to estimate the application performance not only on average, but also in the worst case.

The colloquial **graph center** has a strict definition in the graph-theoretic literature [29]: the set of nodes with minimum eccentricity (equal to the graph radius). The **graph periphery** is the set of nodes with maximum eccentricity (equal to the graph diameter).

Similar to coreness and average distance, we can relate eccentricity to node degree by considering the average eccentricity of  $k$ -degree nodes  $\varepsilon(k)$ .

### 3.2.3 Betweenness

The most commonly used measure of centrality is betweenness [23]. This measure is applicable not only to nodes but also to links [53]. It tries to estimate the potential traffic load on a node or link by counting the number of shortest paths passing through it. Indeed, some variants of betweenness are just called *load* [5].

We use a weighted definition. Let  $\sigma_{ij}$  be the number of shortest paths between nodes  $i$  and  $j$ . Let  $l$  be either the node or link of interest, and let  $\sigma_{ij}(l)$  be the number of shortest paths between  $i$  and  $j$  going through node (or link)  $l$ . Its **betweenness** is then  $B_l = \sum_{ij} \sigma_{ij}(l)/\sigma_{ij}$ . The maximum value of betweenness for both nodes and links is  $n(n-1)$ . To have better comparisons between graphs of different sizes, we can normalize betweenness to 1 by dividing it by its maximum possible value of  $n(n-1)$ .

The simplest approach to calculating node betweenness requires long running times, but we use an efficient algorithm introduced in [5], and modified it to also calculate link betweenness.

As with coreness, average distance, and eccentricity, we consider the relation between betweenness and node degree by calculating the average betweenness of  $k$ -degree nodes  $B(k)$ . For links, we calculate the average link betweenness as a function of degrees of nodes adjacent to a link  $B(k_1, k_2)$ .

Beyond being a measure of centrality trying to estimate the node or link load under uniformly distributed traffic following shortest paths, betweenness is critical for evaluating the accuracy of topology measurements. As shown in [21], this accuracy is higher for graphs with broad betweenness distributions, which is the case with the Internet [11]. As a consequence, the real Internet topology is unlikely to be critically different from the measured one. This observation essentially resolves the “sampling biases” concerns discussed in [38].

### 3.2.4 Spectrum

The adjacency matrix  $\hat{a}$  of a graph has its elements  $a_{ij} = a_{ji} = 1$  if there a link between nodes  $i$  and  $j$ . All other elements are 0. Number  $\lambda$  and vector  $v$  are respectively eigenvalue and eigenvector of  $\hat{a}$  if  $\hat{a}v = \lambda v$ . The **spectrum** of a graph is the set of eigenvalues of its adjacency matrix.

The spectral analysis of graphs is a powerful tool for detailed investigation of graph structure [55, 28], such as discovering clusters of highly interconnected nodes. This technique can reveal the hierarchy of ASs in the Internet [28]. Furthermore, knowing a graph’s spectrum allows one to obtain tight bounds for a wide range of critical graph characteristics, such as distance-related parameters, expansion properties, and values related to separator problems estimating graph resilience under node/link removal [19].

The largest eigenvalues are particularly important. Roughly, graphs with higher values of the largest eigenvalue are of smaller diameter, expand faster, and are harder to break into pieces [19].

### 3.2.5 Entropy

In Section 3.1 we emphasized the level of randomness of graphs. We saw that K-random graphs are “more random” than PK-random graphs, which are “more random” than PKK-random graphs. This question remains: is a unique measure of graph randomness? In other words, can we define a graph entropy?

The rigorous answer to these questions is no. There is no commonly used definition of graph entropy,<sup>2</sup> which is not surprising. Indeed, one can define the graph entropy as the Kolmogorov complexity of its adjacency matrix [10] and one can even use this definition to obtain interesting theoretical bounds for several important graph characteristics [9], but the Kolmogorov complexity is incomputable [40].

We use a different approach inspired by considerations in Section 3.1. Intuitively, classical K-random graphs are the “most random” graphs having the required average degree. At the same time, one can construct [2] the class of K-random graphs via the maximum entropy principle for the node degree distribution.

Indeed, the entropy of a discrete distribution is  $H[P(k)] = -\sum_k P(k) \log_2 P(k)$  [20], and, among the distributions with fixed average  $\bar{k}$ , the distribution with maximum entropy is the binomial distribution, i.e., the one that characterizes the degree distributions in the classical random graphs.

On the other hand, intuitively the most “regular” graphs with fixed average node degree  $\bar{k}$  are  $\bar{k}$ -regular graphs having all nodes of degree  $\bar{k}$ . The entropy of their node degree distribution is at its minimum 0.

These considerations allow us to define the first-order approximation  $\mathcal{H}_K$  to the **graph entropy**  $\mathcal{H}$  as the ratio of the entropy of the node degree distribution in a graph to the entropy of the binomial (or Poisson) distribution with the same average:  $\mathcal{H}_K = H[P(k)]/H[P_K(k)]$ . It is equal to 0 for  $\bar{k}$ -regular graphs and to 1 for K-random graphs. We call  $\mathcal{H}_K$  defined this way the **K-entropy ratio**.

If we have two graphs with the same node degree distributions, then we need the second-order approximation  $\mathcal{H}_{PK}$  to the graph entropy. We can utilize the mutual information [20] of the joint degree distribution  $\mathcal{H}_{PK} = I[P(k_1, k_2)] = H[\tilde{P}(k_1)] + H[\tilde{P}(k_2)] - H[P(k_1, k_2)]$ , where  $\tilde{P}(k)$  is the probability that one of two nodes adjacent to a randomly selected edge is of degree  $k$  (cf. Section 3.1.3). We call  $\mathcal{H}_{PK}$  the **PK-mutual information**. The PK-mutual information captures all the correlations between the two random variables representing degrees of neighboring vertices, not only the linear correlations captured by the assortativity coefficient  $r$ . The PK-mutual information is at its minimum of 0 for uncorrelated PK-random graphs with  $P_{PK}(k_1, k_2) = \tilde{P}(k_1)\tilde{P}(k_2)$  (cf. Section 3.1.3). It attains its maximum on perfectly correlated graphs with  $H[P(k_1, k_2)] = 0$ . In other words,  $\mathcal{H}_{PK}$  tells us how far a given graph is from being PK-random. Intuitively, one can think of PK-mutual information corresponding to the joint degree distribution in the way that K-entropy corresponds to the degree distribution.

We argue that a proper definition of topology graph entropy is a critical component for construction of Internet growth models revealing any fundamental laws of the Internet evolution [15].

<sup>2</sup>The graph entropy defined in information theory (channel and source coding) [32] relates to a graph’s chromatic number [33]: the fewest number of colors necessary to color the vertices of a graph so that no adjacent vertices are of the same color. This definition is not related to graph’s randomness.

## 4 Results

### 4.1 Few notes on BGP tables vs. updates

We first observe that BGP tables and BGP updates are quite similar in terms of vital statistics. In Table 1 we use BGP tables as a baseline to calculate overlap statistics with the other three graph types (BGP updates, skitter, and WHOIS). We show the number of nodes and links present both in BGP tables and the other graph, present only on BGP tables, and present only in the other graph. The BGP-updates column in Table 1 indicates that BGP tables and updates are indeed two similar graphs and not just mostly disjoint graphs having similar numbers of links and edges.

Jumping slightly ahead we are not surprised to notice that all other characteristics of the AS-level topology extracted from BGP updates are remarkably similar to those extracted from BGP tables. Therefore, to make the comparison plots more readable, we do not include statistics for BGP updates; plots with BGP updates are available in the Supplement [11].

The node/edge set differences between BGP tables and skitter<sup>3</sup> or WHOIS in Table 1 are substantial. We directly address these differences in Sections 4.3 and 4.4.

### 4.2 Master comparison of skitter, BGP, and WHOIS

We discuss the calculated topology characteristics in order of their descriptions in Section 3.

**Graph size and average degree.** The first three lines of Table 2 list the number of nodes  $n$  and links  $m$ , and the average degree  $\bar{k} = 2m/n$  in the graphs. BGP sees almost twice as many nodes as skitter. The WHOIS graph is smallest, but its average degree is almost three times larger than that of BGP, and  $\sim 2.5$  times larger than that of skitter. In other words, WHOIS sees substantially more links, both in the absolute ( $m$ ) and relative ( $\bar{k}$ ) senses, than any other data source. We call the order of graphs with increasing average degree  $\bar{k}$ —BGP, skitter, WHOIS—the  **$\bar{k}$ -order**.

**Degree distribution. Figure 1.** Looking at the node degree distribution PDFs and CCDFs, and comparing the observed maximum node degrees  $\bar{k}_{max}$  with those predicted by the clean power laws  $\bar{k}_{max}^{PL}$  (cf. Section 3.1.2), we conclude that skitter follows the power law closest. BGP’s power law is less clean, while the WHOIS graph does not have power laws in its node degree distribution at all! It is not surprising then that, as noticed in [18, 17], the procedure of augmenting the BGP graph with WHOIS links breaks the power law characteristics of the BGP graph.

The WHOIS graph violates the power law by virtue of a significant number of medium degree nodes. As a consequence, its maximum degree is smallest. We can say that the WHOIS graph is least “star-like” (having disproportionately few high-degree nodes), while the skitter graph is most star-like.

As expected, the degree distribution PDFs and CCDFs are in the  $\bar{k}$ -order (BGP < skitter < WHOIS) for a wide range of node degrees.

Finally, we note (cf. Figure 1(a)) that there are fewer 1-degree nodes than 2-degree nodes. The major contributing factor to this phenomenon is the AS number assignment policy [30, 41] that does not require a customer to have an AS number unless it has multiple providers.

**Joint degree distribution. Figures 2 and 3.** The plots in Figure 2 reveal that the PK-random (PLRG) graph with skitter’s node degree distribution has the smallest frequency of tangential links interconnecting medium-degree nodes (the center of Figure 2(a)). The most frequent links are either radial (bottom-right and top-left corners) or low-degree tangential (bottom-left corner).

The real skitter topology is quite different from its PK-random version. The real skitter graph in Figure 2(b) exhibits a relative deficiency of links in the core (top-right corner). These links have either become radial or remained tangential but moved from the core to the medium-degree zone.

Comparing skitter with BGP and WHOIS graphs, we see that skitter has the highest excess of radial links; it is indeed most star-like as suggested by the degree distribution. The WHOIS graph is most tangential, compared to BGP and skitter.

In fact, the WHOIS graph is closest to being PK-random. Recall from Section 3.1.3 that the degree distribution  $P(k)$  fully describes the PK-random graphs. The average neighbor degree of  $k$ -degree nodes  $k_{nn}(k)$  is constant (does not depend on  $k$ ) in the PK-random graphs, and their assortativity coefficient  $r = 0$ . WHOIS is closest to this case based on Figure 3 and values of  $r$  in Table 2.

<sup>3</sup>skitter uses BGP tables to map IP addresses observed in traceroutes to AS numbers. Therefore the number of nodes observed by skitter but not by BGP should be close to 0. That it is not 0 but 1 (AS2277 of degree 2, Ecuonet) is due to the fact that different BGP table dumps were used to construct the BGP table graph and to map an IP address to AS2277 on the day when skitter observed this IP address in its traces.

The skitter graph is on the other extreme: it is the most disassortative (the smallest value of  $r$ ) and its average neighbor degree  $k_{nn}(k)$  has the sharpest decline (the largest value of exponent  $\gamma_{nn}$  of the power-law fit of  $k_{nn}(k)$ ). In other words, the PK-random (PLRG) graph model [1, 53] least accurately describes the skitter topology.

We call the order of graphs with decreasing assortativity coefficient  $r$ —WHOIS, BGP, skitter—the **r-order**.

Finally, we interpret the comparative position of  $k_{nn}(k)$  data points in Figure 3. In the area of low degrees, the most disassortative graph (skitter) is at the top, since the average degree of neighbors of a low-degree node is maximum for the most disassortative graph. In the area of high degrees, the data points are in  $\bar{k}$ -order from bottom to top: WHOIS has the greatest proportion of links contributing to connectivity of neighbors of high-degree nodes.

**Clustering. Figures 4 and 5.** Analyzing clustering, we first detect that the clustering average values are in the  $\bar{k}$ -order, which is expected: more links, more clustering.

Looking at local clustering as a function of node degree  $C(k)$  in Figure 4, we observe an interplay, similar to the case with  $k_{nn}(k)$ , between the  $\bar{k}$ - and  $r$ -orders. Indeed, skitter’s clustering is highest in the area of low degrees because links adjacent to skitter’s low-degree nodes are most radial compared with the other graphs, and hence most likely to lead to high-degree nodes, which are interconnected with high probability as we will see below. In the area of high degrees, the values of clustering for the graph with highest average connectivity (WHOIS) are at the top.

As opposed to  $k_{nn}(k)$ , clustering can help us estimate how well the PKK-random graph model describes the real topology. Figure 5 shows that in the skitter and BGP cases, the local clustering function  $C_{PKK}(k)$  predicted by the PKK-random model follows, albeit shifted down, the form of the observed clustering function  $C(k)$ . To estimate how close these two functions are, we show their mean values  $\bar{C}_{PKK}$  and  $\bar{C}$  in the plots and calculate the ratios of the mean values  $\bar{C}_{PKK}/\bar{C}$  in Table 2. We find that the skitter graph is closest to being PKK-random, while the WHOIS graph is the furthest.

Finally, Figure 5 shows the values of mean local clustering predicted by the PK- and K-random graph models,  $\bar{C}_{PK}$  and  $\bar{C}_K$ . As expected, the PK-random graphs, with constant  $C_{PK}(k)$ , less accurately describe the observed clustering, except in the WHOIS case. Clustering in the K-random graphs is even further away, orders of magnitude smaller than the observed graph. We can estimate accuracy of the K-random model by the ratio  $\bar{C}_K/\bar{C}_{PK}$ . It is highest for WHOIS and lowest for BGP (cf. the entropy discussion below in this section).

**Rich club connectivity. Figure 6.** Rich club connectivity exhibits clean power laws for all four graphs (even for WHOIS) in the area of medium and large  $\rho/n$ . The values of the power-law exponents  $\gamma_{rc}$  in Table 2 result from fitting  $\phi(\rho/n)$  with power laws for 90% of the nodes,  $0.1 \leq \rho/n \leq 1$ .

As expected, the values of  $\phi(\rho/n)$  are in the  $\bar{k}$ -order with WHOIS at the top. However, this pattern breaks in the area of small  $\rho/n$ . Indeed, the WHOIS graph’s top clique size  $n_{clique} = 4$  is surprisingly small. Closer examination of the WHOIS top clique shows that there is only one link missing when adding AS4513 (Globix) with rank  $\rho = 5$ . This link is between AS4513 and AS702 ( $\rho = 1$ , UUNET *EUROPE*). If we assume that this link is present in the WHOIS graph—note that Globix has a link with UUNET *US* (AS701) in all the four graphs,—then the next missing link appears only when adding AS51 (USAREUR) with  $\rho = 15$ . In other words, with the described assumption, the top clique size of WHOIS is expectedly large  $n_{clique} = 14$ .

**Coreness. Figure 7.** Coreness is in the  $\bar{k}$ -order, too, with the WHOIS graph having the largest relative core size and smallest relative fringe size. The coreness of the WHOIS graph is more than three times larger than that of skitter and BGP. WHOIS has a particularly large core size and graph coreness because the  $r$ -order amplifies the  $\bar{k}$ -order here: the WHOIS graph has highest link density (largest  $\bar{k}$ ) and highest concentration of links in the core (largest  $r$ ).

The sparsest graph, the BGP topology, is interesting in that nodes with degree as low as 34 (as high as 7) are in the core (fringe). We also note that the majority of nodes with degrees  $k \gtrsim 100$  are in the core: increasing node degree above 100 does not increase coreness, while for  $k \lesssim 100$  the majority of data points roughly follow power laws.

**Distance. Figures 8 and 9.** All distance distributions have a characteristic Gaussian-like shape, while average distances from  $\bar{k}$ -degree nodes exhibit power laws in the full range of node degrees, even in the WHOIS case.

The skitter graph stands out, and its average distance is smallest, which seems unexpected at first since it has fewer links than the WHOIS graph. One would expect a graph with more links to have lower average distance between nodes. Indeed, adding links to a given graph can only *decrease* the average distance in it. Surprisingly, at first, the WHOIS graph is most richly connected, but its average distance is not lowest. The explanation lies, similar to the cases with average neighbor degree and clustering, in the interplay between the  $\bar{k}$ - and  $r$ -orders.

Indeed, a more disassortative graph has a greater proportion of radial links, shortening the distance from the fringe to the core.<sup>4</sup> The skitter graph has the right balance between the relative number of links and their radiality,

<sup>4</sup>Henceforth, we use terms *fringe* and *core* in a less formal sense, meaning the zones of the graph with low- and high-degree nodes respectively.

to minimize the distance among all other graphs. Compared to skitter, the BGP graph has larger distance because it is sparser (lower  $\bar{k}$ ), and the WHOIS graph has larger distance because it is more assortative (higher  $r$ ).

With these observations, the fact that 62% of AS paths in the skitter graph are 3-hop paths suggests the most frequent pattern: source's AS in the fringe  $\rightarrow$  source's provider AS in the core  $\rightarrow$  destination's provider AS in the core  $\rightarrow$  destination's AS in the fringe.

**Eccentricity. Figures 10 and 11.** Both the eccentricity distribution  $\varepsilon(x)$  and average eccentricity from  $k$ -degree nodes  $\varepsilon(k)$  are similar to their averaged counterparts  $d(x)$  and  $d(k)$ . More interesting are the center and periphery size ratios that are in the  $\bar{k}$ -order, except the center size of the WHOIS graph. In the WHOIS graph, the core consists of only one AS, AS702 (UUNET), which is uniquely positioned to have the minimum eccentricity of 4. If we add to AS702 the nodes having eccentricity of 5, the center becomes much larger, 1109 ASs, and the center size ratio  $n_D/n = 0.1482$  becomes the largest of all four graphs.

**Betweenness. Figures 12 and 14.** The node betweenness is a growing power-law function of node degree, except in the WHOIS graph, which has an excess of medium degree nodes (cf. Figure 1) leading to greater path diversity responsible for lower betweenness values for medium degree nodes.

The contour plots in Figure 14 reveal the important observation that we cannot use link betweenness as a measure of link centrality. Indeed, we see that betweenness of links adjacent to low-degree nodes is not at its minimum. In fact, the betweenness of links adjacent to 1-degree nodes is constant and equal to  $n - 1$ , but similar values of betweenness characterize links elsewhere in the graph, including radial links between low and medium-to-high degree nodes, and tangential links in the zone of medium-to-high degrees. The maximum-betweenness links are in the core as expected, but the minimum-betweenness links are tangential in the medium-to-low degree zone. We can explain the latter observation by the following argument. Let  $i$  and  $j$  be two nodes connected by a minimum-betweenness link  $l$ . The only shortest paths going through  $l$  are those between nodes that are *below*  $i$  and  $j$ , where *below* means further from the core and closer to the fringe. Since the degrees of both  $i$  and  $j$  are small, the numbers of nodes below them (with lower degree) are small, too. Consequently, the number of shortest paths, proportional to the product of the number of nodes below  $i$  and  $j$ , attains its minimum at  $l$ .

We conclude that link betweenness is not a measure of centrality but a measure of some combination of link centrality and radiality.

**Spectrum. Figure 13.** The eigenvalue distributions follow power laws and are in the expected  $\bar{k}$ -order with WHOIS at the top. The WHOIS graph's largest eigenvalue is largest, compared to the other graphs.

**Entropy. Table 3.** First, the maximum value of the K-entropy ratio for the WHOIS graph tells us that this graph is closest to being K-random, compared with the other graphs, while the BGP graph is the least K-random in that respect. This ordering is consistent with the ordering induced by the values of the ratio of the mean local clustering predicted by the K- and PK-random models,  $\bar{C}_K/\bar{C}_{PK}$ , which is an indirect indicator of a graph's proximity to K-randomness.

Second, the values of the PK-mutual information imply that the WHOIS graph is most PK-random, while the skitter graph is least PK-random. This ordering is consistent with the ordering induced by the assortativity coefficient, another direct measure of correlations in  $P(k_1, k_2)$ .

Third, while we do not have a direct measure of PKK-randomness, we recall from the clustering discussion above in this section that the ratio of mean local clustering predicted by the PKK-random model to the observed one,  $\bar{C}_{PKK}/\bar{C}$ , is an indirect measure of PKK-randomness. The skitter (WHOIS) graph is most (least) PKK-random.

We summarize the above three points in Table 3, where we introduce the concept of **dimensions of graph randomness**. One should interpret them as follows. The first dimension, K $\rightarrow$ PK, tells us how well a random graph reproducing just the observed average degree  $\bar{k}$  approximates the observed graph. The second dimension, PK $\rightarrow$ PKK, tells us how well a random graph reproducing the observed degree distribution  $P(k)$  approximates the observed graph. The third dimension, PKK $\rightarrow$ observed, tells us how well a random graph reproducing the observed joint degree distribution  $P(k_1, k_2)$  approximates the observed graph.

The distances in the table (i.e., *closest*, *medium*, and *furthest*) are not absolute values; they are just relative terms inducing ordering among graphs on the same row. Thus it would be wrong to conclude that the WHOIS graph is best described by the K-random model. Indeed, the joint degree distribution contains the most information about the graph, embedding both the degree distribution and the average degree. Therefore, for any given graph, the PKK-random model approximates its topology best.

### 4.3 BGP vs. skitter idiosyncrasies

In contrast to WHOIS and similar to BGP, skitter tries to capture a view of the whole Internet topology. But Table 1 shows significant differences between the two graphs, in terms of (non-)intersecting nodes and links. We seek

to answer the question of where, topologically, these nodes and links are.

We first calculate the node degree distribution of the nodes present only in the BGP graph in Figure 15(a). We detect a skew toward low-degree nodes. The average node degree is 1.86. In other words, nodes seen only in BGP tend to be closer to the fringe, as we would expect.

We also calculate the joint degree distributions for the links present in the intersection of the two graphs and present only in one of the graphs. Figures 15(b)-15(c) show that the tangential links (the diagonals in the plots) in the medium- and high-degree zones are present with similar probabilities both in the intersection of the two topologies and in their difference. The tangential links in the low-degree zone ( $\log_{10}(k) \sim 0$ ) are present with higher probability in the graph difference, consistent with the finding that nodes present only in BGP tend to be of low degrees. There is a shortage of tangential links with slightly higher degrees ( $\log_{10}(k) \sim 1$ ): skitter is more successful in finding such links. The distributions of the radial links (off-diagonal in the plots) do not exhibit any specific pattern.

#### 4.4 BGP vs. WHOIS idiosyncrasies

The WHOIS topology is so different from the other graphs that the following question arises: Can we explain the difference by the fact the WHOIS graph contains only a part of the Internet, namely its European part?

To answer this question we perform the following experiment. We consider the BGP-tables and WHOIS topologies narrowed to the set of nodes present both in BGP tables and WHOIS (cf. Table 1). We find that the induced graphs preserve the full set of the properly normalized topological properties of the original graphs. In Figure 16 we confirm this statement by showing the two most important characteristics: the node degree distribution and average neighbor degree. The full outcome of this experiment is available in the Supplement [11]: the other characteristics exhibit analogous similarities between the full and reduced graphs. Therefore, we can rely on the results of comparison of the full BGP and WHOIS topologies in Section 4.2.

## 5 Conclusions

We calculated the most basic and commonly used statistical characteristics of Internet AS-level topologies extracted from the four most popular sources of AS topology data: skitter measurements, BGP tables, BGP updates, and the RIPE WHOIS database. We compared the four topologies based on a strategically compiled set of graph properties. We found that BGP tables and BGP updates are virtually identical in terms of these properties, while the WHOIS data is substantially different from all the other graphs. The latter difference is essential. It is due to intrinsic differences of the WHOIS data collection mechanics, and it is *not* due to the fact that the WHOIS graph covers primarily European Internet infrastructure.

It might seem unexpected that either the link density ( $\bar{k}$ -order: BGP, skitter, WHOIS), or graph disassortativity ( $r$ -order: WHOIS, BGP, skitter), or their interplay, can explain the relative position of *all* the calculated data curves corresponding to the four topologies. We cannot emphasize enough that the *joint degree distribution*  $P(k_1, k_2)$  contains complete information, both regarding the average degree  $\bar{k}$  responsible for the  $\bar{k}$  order and regarding the assortativity coefficient  $r$  responsible for the  $r$ -order. In other words, the following picture of dependencies between topological characteristics emerges. The joint degree distribution defines both the  $\bar{k}$ - and  $r$ -orders, which, in turn, explain the relative position of *all other* characteristic data curves we calculate for the four compared topologies.

While  $P(k_1, k_2)$  contains the most complete topological information about the graph, we also introduce the concept of *dimensions of graph randomness* to estimate how well the random graph models that reproduce the average node degree  $\bar{k}$  (K-random), degree distribution  $P(k)$  (PK-random), or joint degree distribution  $P(k_1, k_2)$  (PKK-random) approximate the observed topologies. We find the PKK-random model describes the skitter graph most accurately. If we constrain ourselves to *less accurate* graph descriptions (PK- or K-random), then they best describe the WHOIS graph. The latter fact implies that the PLRG random graph model and topology generators based on it cannot accurately capture the important properties of the skitter or BGP graphs. These observations have strong implications for the development of new and credibility of existing network topology generators.

Two important questions remain. The first is: can we explain the  $\bar{k}$ - and  $r$ -orders by the specifics of the data sources? It turns out that the explanation of WHOIS's highest  $\bar{k}$  and  $r$  is not complex. Indeed, both skitter and BGP are *traceroute-like* explorations of the network topology, meaning that we can closely approximate the graphs extracted from skitter and BGP by a union of spanning trees rooted at, respectively, skitter monitors or BGP data collection points. As such, both skitter and BGP fail to detect many tangential links since these links do not lie on any shortest path rooted at a vantage point in the core. WHOIS, on the other hand, contains those links as directly

attached to sources of WHOIS records (values of `aut-num` fields). The abundance of these tangential medium-degree links increase both the average degree  $\bar{k}$  and assortativity coefficient  $r$ .

The explanation of the  $\bar{k}$ - and  $r$ -orders for skitter and BGP is less straightforward. We saw in Section 4.3 that nodes that are in BGP but not in skitter are mostly of low degree. The skitter graph does not contain these low-degree ASs because they are so small that skitter’s list of traceroute destinations fails to contain any replying IP address in the IP address blocks these ASs advertise. Since the BGP graph has relatively more low-degree nodes, its average degree is lower. We also saw in Section 4.3 that most links present only in BGP are tangential links between low-degree ASs. The majority of such links should connect the low-degree ASs present only in BGP to their secondary (backup) small, low-degree providers, while their primary provider is large and high-degree. Even if skitter detects a low-degree AS having such a small backup provider, skitter is still unlikely to detect its backup link since traceroutes follow the primary path via the large provider. Being tangential, these backup links increase the BGP graph’s assortative coefficient.

The second question is: which observed topology describes the real Internet AS topology most accurately? We can try to approach this question from several angles.

- As discussed in Section 2, skitter measurements occur via the data plane and therefore are most likely to directly reflect topology seen by Internet traffic. The WHOIS data is on the other extreme: it is manually maintained with no requirements for timely updates, and is therefore most likely to contain stale or inaccurate information. The authors of [51] attack this problem by developing a set of data-cleaning techniques, but application of these techniques does not drastically change the statistical characteristics of the WHOIS-derived graph [51].
- Of course, the skitter data is not pruned of errors and noise either. However, scanning through Figures 1-14 we notice that the skitter statistics appear least noisy, in the sense that we can most accurately approximate them by curves. Somewhat surprisingly, the noisiest dataset is not WHOIS, but BGP.
- Recent work [21] increases the credibility of traceroute-like sampling (e.g., skitter or BGP) of realistic topologies. The WHOIS data collection process is not traceroute-like, and the drastic difference between the WHOIS-derived topology and the skitter or BGP topologies is in conflict with increased credibility of traceroute-like explorations. A rigorous resolution of this conflict calls for additional analytical and experimental efforts.
- Given the coarsest and most basic characteristic of connectivity, the  $\bar{k}$ -order, we can say that BGP suffers from link under-sampling, while WHOIS suffers from link over-sampling. It is tempting to say that the truth lies somewhere in between, approximately where the skitter graph lies.

The most cautious answer to the question of which data source is closest to reality would be that there is not one but *three* sources of vital statistics of Internet AS-level topology: skitter, BGP, and WHOIS. They are all different, with WHOIS being most different from the others. We refer those who are satisfied with only one of them—with skitter, for example—to Table 3: a random graph generator reproducing the observed joint degree distribution should accurately capture a wide range of critical characteristics of the AS-level topology observed by skitter.

In the Supplement [11] we make available both the original Internet AS-level graphs and the calculated statistics.

## 6 Acknowledgements

We thank Ulrik Brandes for sharing his betweenness code with us and Andre Broido for answering our questions.

Support for this work was provided by NSF CNS-0434996, NCS ANI-0221172, Cisco’s University Research program, and other CAIDA members.

## References

- [1] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the 32<sup>nd</sup> Annual ACM Symposium on Theory of Computing (STOC)*, pages 171–180. ACM Press, 2000.
- [2] M. Bauer and D. Bernard. Maximal entropy random networks with given degree distribution. <http://arxiv.org/abs/cond-mat/0206150>.
- [3] B. Bollobás. *Random Graphs*. Academic Press, New York, 1985.

- [4] B. Bollobás and O. Riordan. Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks*, Berlin, 2002. Wiley-VCH.
- [5] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.
- [6] A. Broido and kc claffy. Internet topology: Connectivity of IP graphs. In *SPIE International Symposium on Convergence of IT and Communication*, 2001.
- [7] A. Broido, E. Nemeth, and kc claffy. Internet expansion, refinement, and churn. *European Transactions on Telecommunications*, 13(1):33–51, 2002.
- [8] T. Bu and D. Towsley. On distinguishing between Internet power law topology generators. In *IEEE INFOCOM*, 2002.
- [9] H. Buhrman, J.-H. Hoepman, and P. Vitányi. Space-efficient routing tables for almost all networks and the incompressibility method. *SIAM Journal on Computing*, 28(4):1414–1432, 1999.
- [10] H. Buhrman, M. Li, John J. Tromp, and P. Vitányi. Kolmogorov random graphs and the incompressibility method. *SIAM Journal on Computing*, 29(2):590–599, 2000.
- [11] CAIDA. Comparative analysis of the Internet AS-level topologies extracted from different data sources: Data page. [http://www.caida.org/analysis/topology/as\\_topo\\_comparisons/](http://www.caida.org/analysis/topology/as_topo_comparisons/).
- [12] CAIDA. Macroscopic topology AS adjacencies. [http://www.caida.org/tools/measurement/skitter/as\\_adjacencies.xml](http://www.caida.org/tools/measurement/skitter/as_adjacencies.xml).
- [13] CAIDA. Macroscopic topology measurements. Research Project. <http://www.caida.org/analysis/topology/macroscopic/>.
- [14] CAIDA. rv2atoms. <http://www.caida.org/projects/routing/atoms/download/>.
- [15] CAIDA. Toward mathematically rigorous next-generation routing protocols for realistic network topologies. Research Project. <http://www.caida.org/projects/nets-nr/>.
- [16] CAIDA. Visualizing Internet topology at a macroscopic scale. [http://www.caida.org/analysis/topology/as\\_core\\_network/](http://www.caida.org/analysis/topology/as_core_network/).
- [17] H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. Towards capturing representative AS-level Internet topologies. *Computer Networks Journal*, 44:737–755, April 2004.
- [18] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. The origin of power laws in Internet topologies revisited. In *IEEE INFOCOM*, 2002.
- [19] F. K. R. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997.
- [20] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.
- [21] L. Dall’Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. A statistical approach to the traceroute-like exploration of networks: Theory and simulations. <http://arxiv.org/abs/cond-mat/0406404>.
- [22] S. N. Dorogovtsev. Clustering of correlated networks. *Physical Review E*, 69:027104, 2004.
- [23] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.
- [24] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Metric structure of random networks. *Nuclear Physics B*, 653(3):307–422, 2003.
- [25] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.
- [26] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *ACM SIGCOMM*, pages 251–262, 1999.
- [27] M. Gaertler and M. Patrignani. Dynamic analysis of the Autonomous System graph. In *IPS*, 2004.
- [28] C. Gkantsidis, M. Mihail, and E. Zegura. Spectral analysis of Internet topologies. In *IEEE INFOCOM*, 2003.
- [29] F. Harary. *Graph Theory*. Addison-Wesley, Reading, MA, 1994.
- [30] J. Hawkinson and T. Bates. *Guidelines for Creation, Selection, and Registration of an Autonomous System (AS)*. IETF, RFC 1930, 1996.
- [31] Y. Hyun, A. Broido, and kc claffy. Traceroute and BGP AS path incongruities. In *Cooperative Association for Internet Data Analysis (CAIDA)*, 2003. <http://www.caida.org/outreach/papers/2003/ASP/>.
- [32] J. Körner. Coding of an information source having ambiguous alphabet and the entropy of graphs. In *Transactions of the 6<sup>th</sup> Prague Conference on Information Theory*, pages 411–425, 1973.
- [33] J. Körner and A. Orłitsky. Zero-error information theory. *IEEE Transactions on Information Theory*, 44(6):2207–2229, 1998.
- [34] S. Jaiswal, A. L. Rosenberg, and D. Towsley. Comparing the structure of power-law graphs and the Internet AS graph. In *IEEE ICNP*, 2004.
- [35] kc claffy, T. E. Monk, and D. McRobb. Internet tomography. *Nature*, January 1999. <http://www.caida.org/>

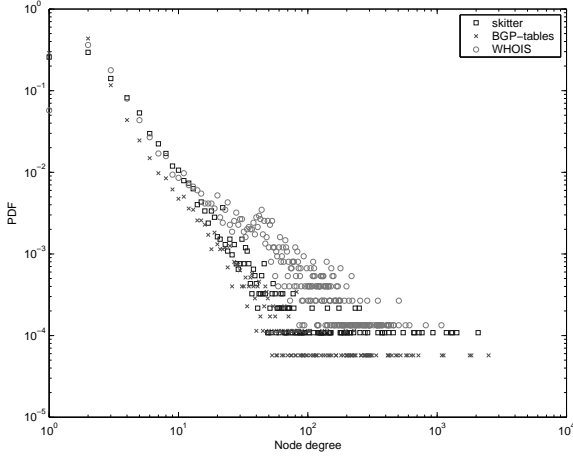
tools/measurement/skitter/.

- [36] D. Krioukov, K. Fall, and X. Yang. Compact routing on Internet-like graphs. Technical Report IRB-TR-03-010, Intel Research, 2003.
- [37] D. Krioukov, K. Fall, and X. Yang. Compact routing on Internet-like graphs. In *IEEE INFOCOM*, 2004.
- [38] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, 2003.
- [39] L. Li, D. Alderson, W. Willinger, and J. Doyle. A first-principles approach to understanding the Internets router-level topology. In *ACM SIGCOMM*, 2004.
- [40] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 1997.
- [41] M. Kühne, N. Nimpuno, and S. Wilmot. *Autonomous System (AS) Number Assignment Policies and Procedures*. RIPE, ripe-263, 2003.
- [42] Z. M. Mao, D. Johnson, J. Rexford, J. Wang, and R. Katz. Scalable and accurate identification of AS-level forwarding paths. In *IEEE INFOCOM*, 2004.
- [43] Z. M. Mao, J. Rexford, J. Wang, and R. H. Katz. Towards an accurate AS-level traceroute tool. In *ACM SIGCOMM*, 2003.
- [44] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–179, 1995.
- [45] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7:295–305, 1998.
- [46] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.
- [47] D. Peleg. *Distributed Computing: A Locality-Sensitive Approach*. SIAM, Philadelphia, PA, 2000.
- [48] Internet Routing Registries. <http://www.irr.net/>.
- [49] Y. Rekhter and T. Li. *A Border Gateway Protocol 4 (BGP-4)*. IETF, RFC 1771, 1995.
- [50] University of Oregon Route Views Project. <http://www.routeviews.org/>.
- [51] G. Siganos and M. Faloutsos. Analyzing BGP policies: Methodology and tool. In *IEEE INFOCOM*, 2004.
- [52] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos. Power-laws and the AS-level Internet topology. *ACM/IEEE Transactions on Networking*, 11(4):514–524, 2003.
- [53] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. Network topology generators: Degree-based vs. structural. In *ACM SIGCOMM*, pages 147–159, 2002.
- [54] traceroute. <http://www.traceroute.org/#source%20code>.
- [55] D. Vukadinović, P. Huang, and T. Erlebach. A spectral analysis of the Internet topology. Technical Report TIK-NR. 118, ETH, 2001.
- [56] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. Scaling phenomena in the Internet: Critically examining criticality. *PNAS*, 99(Suppl. 1):2573–2580, 2002.
- [57] X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. An analysis of BGP multiple origin AS (MOAS) conflicts. In *ACM SIGCOMM IMW*, 2001.
- [58] S. Zhou and R. J. Mondragon. Accurately modeling the Internet topology. <http://arxiv.org/abs/cs.NI/0402011>.

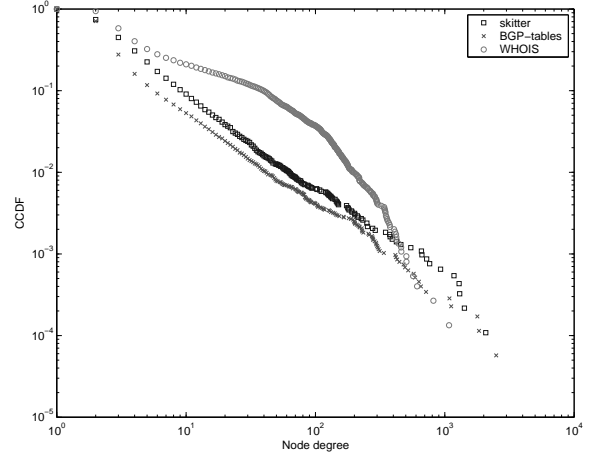
## Tables and Figures

Table 1: **BGP tables compared with BGP updates, skitter, and WHOIS.** The BGP-tables topology is graph  $G_A(V_A, E_A)$  with the sets of nodes  $V_A$  and edges  $E_A$ . Graph  $G_B(V_B, E_B)$  is one the other graphs listed in the first row.

	BGP updates	skitter	WHOIS
Number of nodes in both $G_A$ and $G_B$ ( $V_A \cap V_B$ )	17,349	9,203	5,583
Number of nodes in $G_A$ but not in $G_B$ ( $V_A \setminus V_B$ )	97	8,243	11,863
Number of nodes in $G_B$ but not in $G_A$ ( $V_B \setminus V_A$ )	68	1	1,902
Number of edges in both $G_A$ and $G_B$ ( $E_A \cap E_B$ )	38,543	17,407	12,335
Number of edges in $G_A$ but not in $G_B$ ( $E_A \setminus E_B$ )	2,262	23,398	28,470
Number of edges in $G_B$ but not in $G_A$ ( $E_B \setminus E_A$ )	3,941	11,552	44,614

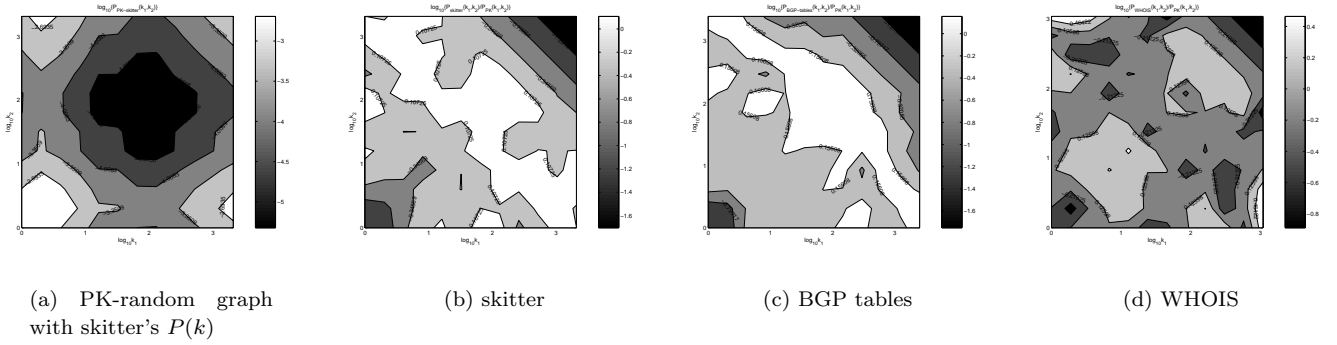


(a) PDF



(b) CCDF

Figure 1: Node degree distribution  $P(k)$ .



(a) PK-random graph with skitter's  $P(k)$

(b) skitter

(c) BGP tables

(d) WHOIS

Figure 2: **Joint degree distribution  $P(k_1, k_2)$ .** **a)** The contour plot of the logarithm of the joint degree distribution  $P_{PK}(k_1, k_2)$  for a PK-random graph (cf. Section 3.1.3), given the skitter degree distribution  $P(k)$ . **b)** The logarithm of the ratio of  $P(k_1, k_2)$  observed in the real skitter graph to  $P_{PK}(k_1, k_2)$  above. **c,d)** The diagrams, analogous to (b), for BGP and WHOIS. Some asymmetry of the diagrams is due to interpolation and rounding algorithms in MATLAB. The *scatter* plots in the Supplement [11] are symmetric.

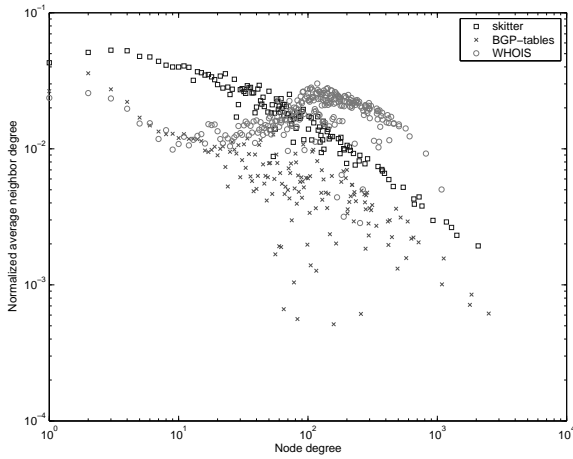


Figure 3: Normalized average neighbor degree  $k_{nn}(k)/(n-1)$ .

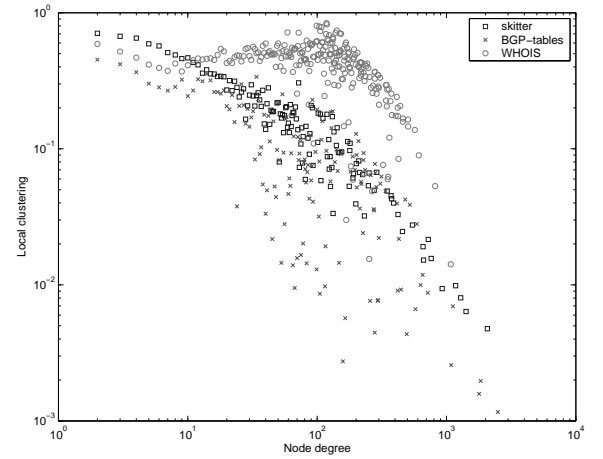
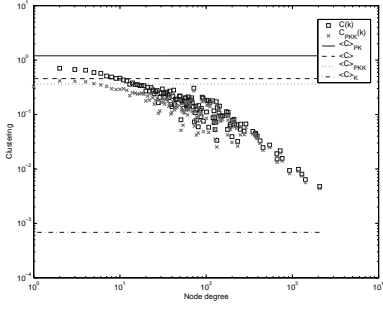
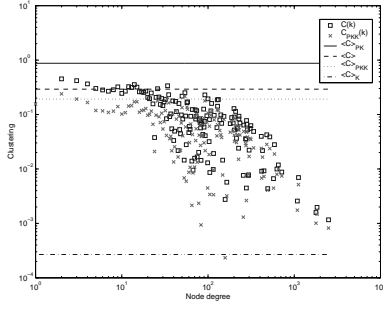


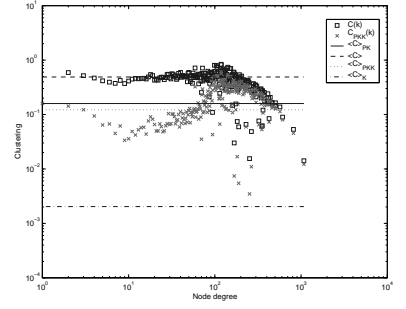
Figure 4: Local clustering  $C(k)$ .



(a) skitter



(b) BGP tables



(c) WHOIS

Figure 5: **Local clustering vs. graph randomness.** Squares show local clustering observed in the real topology. The dashed line is its mean value. Crosses show local clustering predicted by the PKK-random graph model  $C_{PKK}(k)$ . The dotted line is its mean value. The solid and dash-dotted lines are constant clusterings predicted by the PK- and K-random graph models,  $C_{PK}(k)$  and  $C_K(k)$  from Section 3.1.4. Notation  $\langle C \rangle$  in the legends means the same as  $\bar{C}$  in the text.

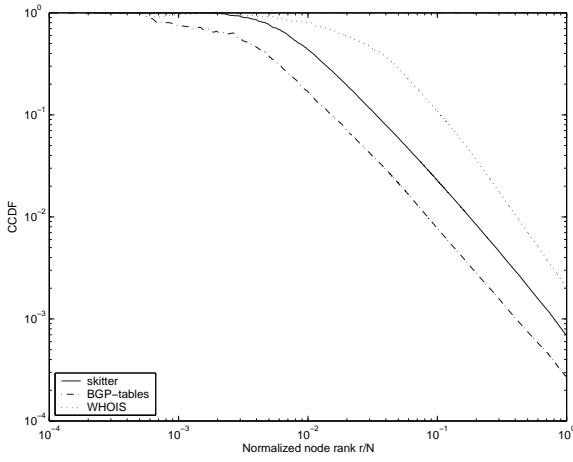


Figure 6: Rich club connectivity  $\phi(\rho/n)$ .

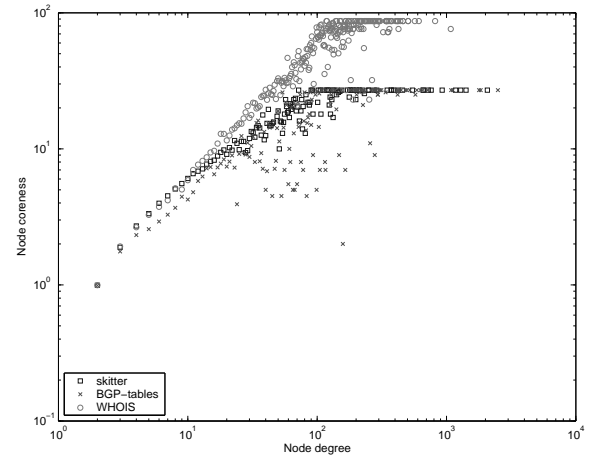


Figure 7: Average coarsness of  $k$ -degree nodes  $\kappa(k)$ .

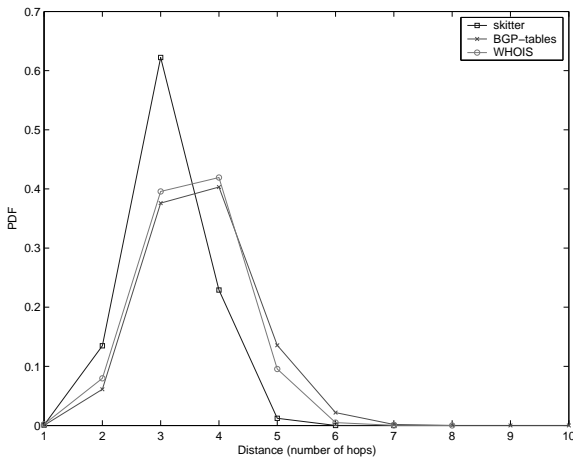


Figure 8: Distance distribution  $d(x)$ .

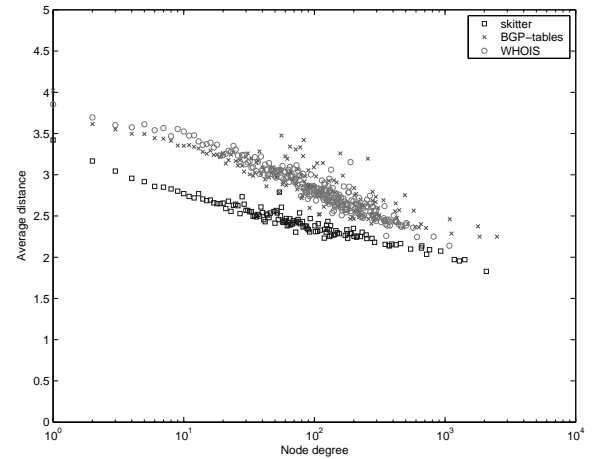


Figure 9: Average distance from  $k$ -degree nodes  $d(k)$ .

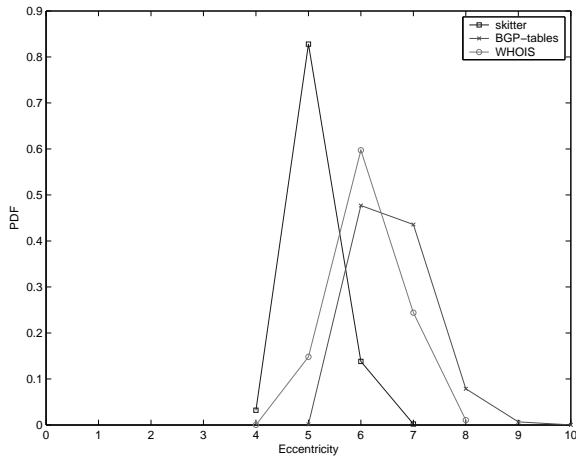


Figure 10: Eccentricity distribution  $\varepsilon(x)$ .

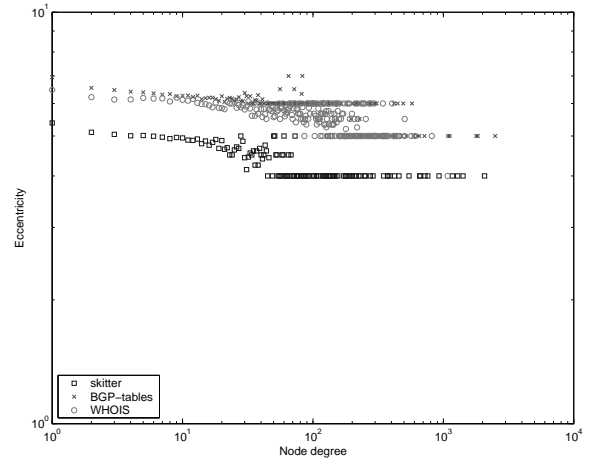


Figure 11: Average eccentricity of  $k$ -degree nodes  $\varepsilon(k)$ .

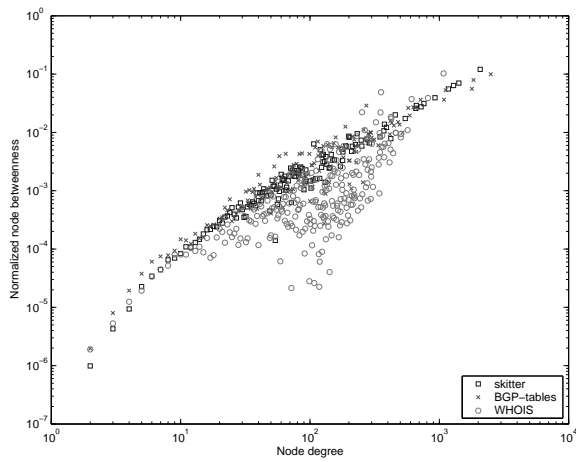


Figure 12: Normalized node betweenness  $B(k)/(n(n-1))$ .

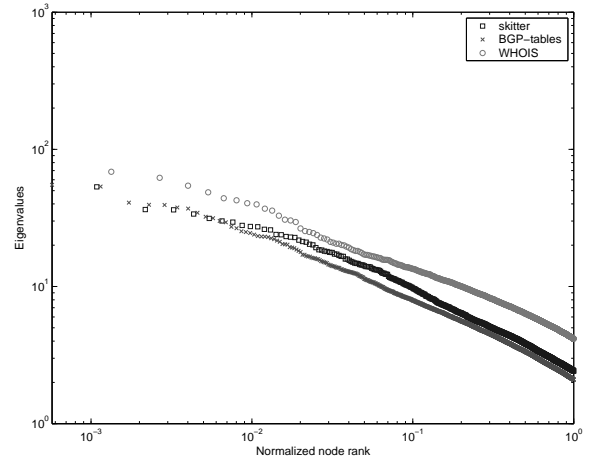


Figure 13: **Spectrum.** Absolute values of top 10% of eigenvalues ordered by their normalized rank: the absolute value divided by the total number of eigenvalues calculated for a given graph.

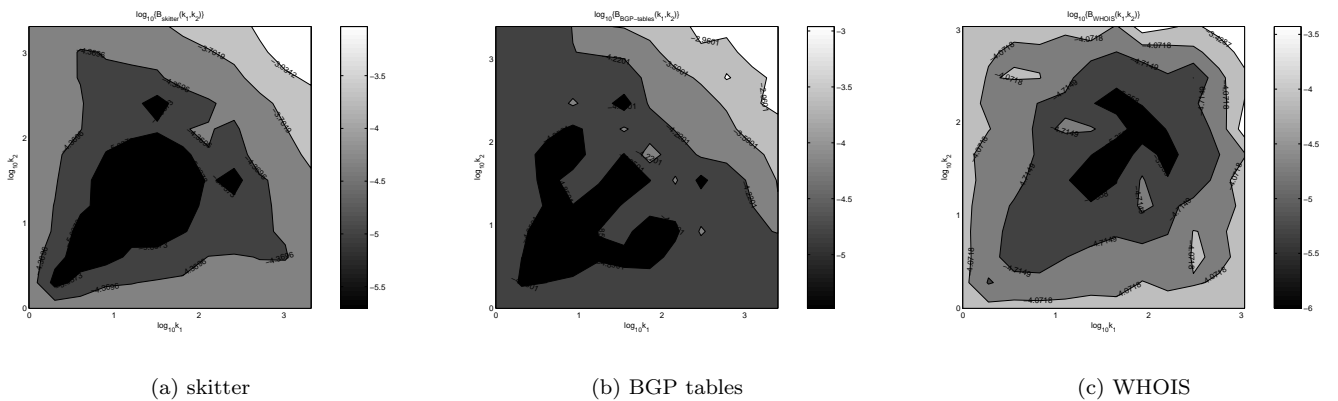


Figure 14: Normalized link betweenness  $B(k_1, k_2)/(n(n-1))$ . The contour plots of logarithms of normalized link betweenness as functions of logarithms of degrees of adjacent nodes.

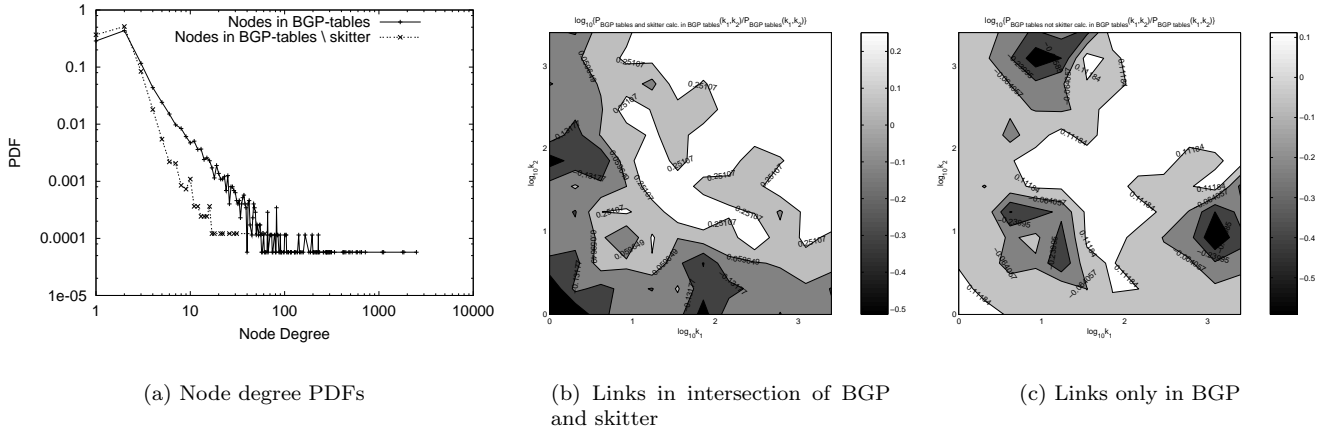


Figure 15: **BGP tables vs. skitter.** **a)** Degree distributions of nodes present in the full BGP-tables graph and in the BGP-tables graph only, not in skitter. **b,c)** The contour plots of logarithms of the ratios  $P_{BGP \cap skitter}(k_1, k_2)/P_{BGP}(k_1, k_2)$  and  $P_{BGP \setminus skitter}(k_1, k_2)/P_{BGP}(k_1, k_2)$ , where  $P_{BGP}(k_1, k_2)$  is the joint degree distribution in the full BGP-tables graph,  $P_{BGP \cap skitter}(k_1, k_2)$  is the joint degree distribution of links present in the intersection of the BGP-tables and skitter graphs, and  $P_{BGP \setminus skitter}(k_1, k_2)$  is the joint degree distribution of links present only in the BGP-tables graph. The node degrees are calculated in the full BGP-tables graph. The skitter counterparts of (b) and (c) are available in the Supplement [11].

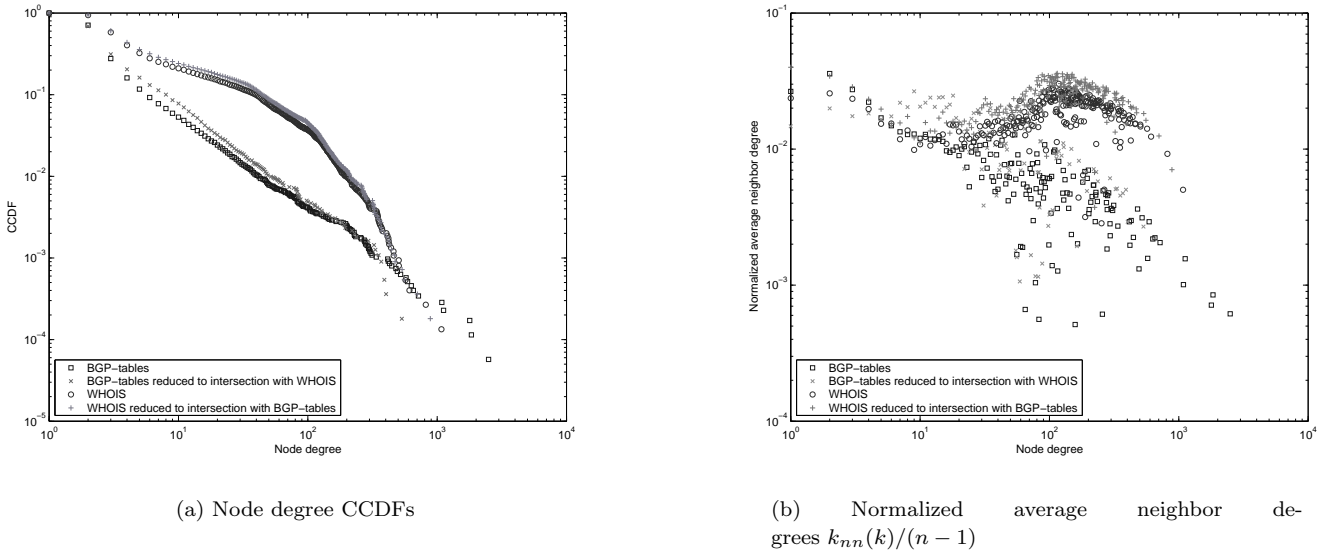


Figure 16: **BGP tables vs. WHOIS.**

Table 2: Summary statistics.

		skitter	BGP tables	BGP updates	WHOIS
Average degree	Number of nodes ( $n$ )	9,204	17,446	17,417	7,485
	Number of edges ( $m$ )	28,959	40,805	42,484	56,949
	Avg node degree ( $k$ )	6.2927	4.6779	4.8785	15.2168
Degree distr	Max node degree ( $k_{max}$ )	2,070	2,498	2,627	1,079
	Max degree ratio ( $k_{max}/(n-1)$ )	0.2249	0.1432	0.1508	0.1442
	Power-law max degree ( $k_{max}^{PL}$ )	1,448	4,546	4,331	-
	Exponent of $P(k)$ ( $-\gamma$ )	2.2541	2.1597	2.1662	-
Joint degree distr	Avg neighbor degree ( $k_{nn}/(n-1)$ )	0.0469	0.0289	0.0303	0.0212
	Max neighbor degree ( $k_{nn}^{max}/(n-1)$ )	0.0530	0.0359	0.0375	0.0302
	Exponent of $k_{nn}(k)$ ( $-\gamma_{nn}$ )	1.4886	1.4452	1.4548	-
	Assortative coefficient ( $r$ )	-0.2356	-0.1936	-0.1930	-0.0424
Clustering	Mean clustering ( $\bar{C}$ )	0.4567	0.2918	0.3280	0.4887
	PKK/measured ratio ( $\bar{C}_{PKK}/\bar{C}$ )	0.7925	0.6578	0.5918	0.2505
	K/PK ratio ( $\bar{C}_K/\bar{C}_{PK}$ )	$5.71 \cdot 10^{-4}$	$3.06 \cdot 10^{-4}$	$3.14 \cdot 10^{-4}$	0.0128
	Clustering coefficient ( $C$ )	0.0258	0.0155	0.0157	0.3062
	Exponent of $C(k)$ ( $-\gamma_C$ )	0.3324	0.3415	0.3421	-
Rich club	Top clique size ( $n_{clique}$ )	16	9	7	4
	Exponent of $\phi(\rho/n)$ ( $-\gamma_{rc}$ )	1.4814	1.4488	1.4549	1.6851
Coreness	Min node coreness ( $\kappa_{min}$ )	0	0	0	0
	Avg node coreness ( $\bar{\kappa}$ )	2.2286	1.4146	1.5212	7.6546
	Max node coreness ( $\kappa_{max}$ )	27	27	27	87
	Core size ratio ( $n_{core}/n$ )	0.0051	0.0030	0.0037	0.0171
	Min degree in core ( $k_{core}^{min}$ )	68	34	31	99
	Fringe size ratio ( $n_{fringe}/n$ )	0.2670	0.2916	0.2542	0.0591
	Max degree in fringe ( $k_{fringe}^{max}$ )	5	7	7	4
	Exponent of $\kappa(k)$ ( $\gamma_{\kappa}$ )	0.6751	0.5821	0.6386	1.0692
Distance	Avg distance ( $\bar{d}$ )	3.1180	3.6856	3.6454	3.5433
	Std deviation of distance ( $\sigma$ )	0.6348	0.8745	0.8610	0.8015
	Exponent of $d(k)$ ( $-\gamma_d$ )	0.0710	0.0732	0.0732	0.0895
Eccentricity	Graph radius ( $R, \varepsilon_{min}$ )	4	5	5	4
	Avg eccentricity ( $\bar{\varepsilon}$ )	5.1102	6.6127	6.4358	6.1166
	Graph diameter ( $D, \varepsilon_{max}$ )	7	10	9	8
	Center size ratio ( $n_R/n$ )	0.0320	0.0014	0.0040	$1.33 \cdot 10^{-4}$
	Min degree in center ( $k_R^{min}$ )	4	188	4	1,079
	Periphery size ratio ( $n_D/n$ )	0.0021	$1.71 \cdot 10^{-4}$	0.0014	0.0106
	Max degree in periphery ( $k_D^{max}$ )	1	1	2	6
Betweenness	Max node betweenness ( $B_{node}^{max}/n(n-1)$ )	0.1205	0.0991	0.1011	0.1027
	Avg node betweenness ( $\bar{B}_{node}/n(n-1)$ )	$1.14 \cdot 10^{-4}$	$7.69 \cdot 10^{-5}$	$7.59 \cdot 10^{-5}$	$1.70 \cdot 10^{-4}$
	Exponent of $B(k)$ ( $\gamma_B$ )	1.3470	1.1668	1.3080	-
	Min edge betweenness ( $B_{edge}^{min}/n(n-1)$ )	$1.18 \cdot 10^{-8}$	$3.28 \cdot 10^{-9}$	$3.29 \cdot 10^{-9}$	$1.78 \cdot 10^{-8}$
	Avg edge betweenness ( $\bar{B}_{edge}/n(n-1)$ )	$5.37 \cdot 10^{-5}$	$4.51 \cdot 10^{-5}$	$4.29 \cdot 10^{-5}$	$3.10 \cdot 10^{-5}$
	Max edge betweenness ( $B_{edge}^{max}/n(n-1)$ )	0.0043	0.0062	0.0049	0.0064
Spectrum	Largest eigenvalue	79.5338	73.0592	74.8988	150.8588
Entropy	K-entropy ratio ( $\mathcal{H}_K$ )	0.9360	0.8171	0.8222	0.9981
	PK-mutual information ( $\mathcal{H}_{PK}$ )	0.4269	0.5999	0.5734	0.6212

Table 3: Dimensionality of graph randomness.

Dimensions	How close to being...	Measures	skitter	BGP	WHOIS
K→PK	K-random	$\mathcal{H}_K, \bar{C}_K/\bar{C}_{PK}$	medium	furthest	closest
PK→PKK	PK-random	$\mathcal{H}_{PK}, r$	furthest	medium	closest
PKK→observed	PKK-random	$\bar{C}_{PKK}/\bar{C}$	closest	medium	furthest