IEICE TRANS. COMMUN., VOL.E96–B, NO.3 MARCH 2013

LETTER

Revisiting the Rich Club of the Internet AS-Level Topology*

Yangyang WANG[†a)], *Nonmember*, **Jun BI**[†b)], *Member*, *and* **Jianping WU**[†], *Nonmember*

SUMMARY We evaluate the rich-club property of the Internet topology at the autonomous system (AS) level by comparing the Internet AS graphs of traceroute and BGP, and the synthetic graphs of PFP model. The results indicate that, for rich-club coefficient, PFP model can exactly match traceroute AS graphs in the early years around 2002, but it has significantly deviated from the grown AS graphs since about 2010.

key words: *Internet topology, topology model, topology evolution*

1. Introduction

Understanding and modeling the Internet topology at the autonomous system (AS) level can benefit network protocol design and evaluation. The rich-club property is an important property of the Internet AS topology. It describes a topological nature that the "rich" AS nodes are densely connected to each other. The richness of a node is measured by its node degree, i.e., the number of nearest neighbors of this node. And the rich-club property can be quantitatively defined by the rich-club coefficient [1]. To calculate the rich-club coefficient of a network, nodes are sorted in decreasing order of node degrees. A group of nodes of the same node degree can be sorted arbitrarily in that group. The position of a node in the resulting ordered sequence is referred to as its rank. The rich-club coefficient $\phi(r)$ of the top r nodes is defined as

$$\phi(r) = \frac{2e}{r(r-1)} \tag{1}$$

It is the ratio of the actual number of links e to the maximum possible number of links among the r nodes. If the entire graph has N nodes, the rich-club coefficient can be normalized as $\phi(r/N)$. The work in [1] shows that the Internet AS topology has a notable rich-club property with $\phi(1\%) > 32\%$. Some studies [2]–[4] debated and confirmed the presence of rich-club property in various complex networks, including the Internet AS-level topology. A few Internet topology models can produce topologies with a notable rich-club feature, in which the Positive-Feedback Preference (PFP) model is more accurate than other models in terms of a number of topological properties [5], [6]. In this letter, we revisit the rich-club property and show the variations and differences of the measured AS graphs and PFP model as the Internet grows.

Manuscript received January 16, 2012.
Manuscript revised September 24, 2012.

[†]The authors are with the Network Research Center, Tsinghua University, 100084, Beijing, China.

[*]This work was supported by National Natural Science Foundation of China (No. 61073172). and 973 Program of China (No. 2009CB320501)

a) E-mail: wyystar@gmail.com
b) E-mail: junbi@tsinghua.edu.cn

DOI: 10.1587/transcom.E96.B.900

2. PFP Model

The PFP model, described in [5], [7], generates network topologies using a nonlinear preference probability and an interactive growth mechanism. It starts with a small random graph. The nodes in the existing graph are called old nodes. Then, at each step, 1) with probability $p \in [0, 1]$, a new node is added and connected to an old node, named *host* node, and two new links are established between the host node and two other old nodes; 2) with probability $1 - p$, a new node is connected to two host nodes, and one of the two host nodes establishes a link to another old node in the graph. In this process, the links between old nodes, referred to as internal links, increase the connectivity among the rich nodes. A new node is connected to a node of degree k with the nonlinear preference probability

$$\Pi(k) = \frac{k^{1+\delta \log_{10} k}}{\sum_j k_j^{1+\delta \log_{10} k_j}} \tag{2}$$

In this formula, the node degree k is used as the positive feedback of the preference probability. It has been studied that the parameters $\delta = 0.048$ and $p = 0.4$ produce the best result that most closely matches the Internet AS topology [6]. Therefore, in our simulations, PFP topologies are generated with $\delta = 0.048$ and $p = 0.4$.

3. Topology Dataset

Our AS topology datasets include BGP data, traceroute measurements, and PFP model generated data. Because of the incompleteness of and faults in the publicly available measurement data, we collected AS topology data from different sources to give multiple witnesses of different views for evaluation.

BGP: Border Gateway Protocol (BGP) is the routing protocol connecting ASes in the Internet. BGP routing tables contain AS path attribute that can be used to construct AS graphs. We created AS graph datasets using the BGP data from RIPE RIS [8] and RouteViews [9] in the first day of a specific month. These datasets are named bgp0102, bgp1201, and bgp0304. The name 'bgp1210' means it is

from the BGP data in the first day of December 2010. The other names follow the same pattern. The Internet topology project of UCLA [10] collects AS links from a wider range of BGP data feeds, including router servers, looking glasses, etc., starting from the year 2004. The dataset ucla1210 contains accumulated AS links observed during the entire month of December 2010. The public BGP data covers the majority of AS links, but is not complete. Many links of peer-peer business relationship are missing in the public BGP data [11].

Traceroute: The traceroute AS data is partly got from the Skitter project of CAIDA, which is updated as the Ark project after 2008 [12]. We downloaded all daily snapshots in a specific month and merged them into one dataset file. Similar to the work in [13], the AS links derived from multi-origin ASes, AS-sets, and private ASes are filtered in our datasets. This part of datasets include ski0304, ski0102 and ark1210. The name 'ski0304' means it is from Skitter's data in March, 2004. The other names follow the same pattern. Another part of traceroute AS data is sidewalk [14] that provides accumulated AS links observed from Dec. 2007 to Sept. 2008. Traceroute measurement also cannot scan every link in the Internet, and the conversion from tranceroute IP sequences to AS graphs needs to map IP addresses to AS numbers using BGP routing tables, which proved to be error-prone with false AS links [15]. The work in [15] shows that sidewalk is less inaccurate than Skitter and Ark through its filtering of false links.

PFP model: Similar to the previous work in [6], we used PFP model to generate ten topology instances of the same number of nodes as the measured AS topology, starting from different small random graphs. And the topological metrics are averaged over the ten instances. The datasets from PFP model include PFP_ski0304, PFP_ski0102, PFP_ark1201 for traceroute data, and PFP_bgp0102, PFP_bgp1201 for BGP data, and PFP_(ski+bgp)0102, PFP_(ark+bgp)1201 for the merged data of traceroute and BGP data of the same timestamp.

4. Metrics of AS Graphs

To show the difference of these topology datasets, we list a few statistical metrics about AS graphs in Table 1 and Table 2. And some metrics are described briefly in the following:

- **Average degree**. It is the ratio of the sum of all node degrees to the total number of nodes. Suppose that a graph has E links and N nodes, then the average degree is $2E/N$.
- **Average shortest path length**. The shortest path length is the number of links in the shortest path between two nodes. The average shortest path length is the shortest path length averaged over all pairs of nodes.
- **Assortative coefficient**. This metric is defined as [16]

$$\alpha = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2} \quad (3)$$

  In the above equation, M represents the number of links in the graph, and the j_i and k_i denote the degrees of the two end nodes of the ith link. It reflects the correlation of node degrees on the links. The Internet AS topology has $\alpha < 0$, which implies it is a disassortative network where high degree nodes prefer to connect with the nodes of low degree.
- **Rich-club clique**. It is the number of nodes in the maximum rich-club that has a rich-club coefficient $\phi = 1.0$.
- **Clustering coefficient**. The local clustering coefficient

Table 1 Comparison of PFP and measured datasets in the early years about 2002.

Metrics	ski0102	PFP	bgp0102	PFP	(bgp+ski)0102	PFP	ski0304	PFP
number of nodes, N	7168	7168	12503	12503	12567	12567	9204	9204
number of links, L	23246	21431	26535	37390	37332	37525	28959	27492
maximum degree, K	1763	1942	2553	2737	3088	3928	2070	2714
average degree, $\bar{k}$	6.49	5.98	4.24	5.98	5.94	5.97	6.29	5.97
average shortest path, $\bar{l}$	3.0	3.11	3.62	3.11	3.11	2.99	3.13	3.03
assortative coefficient, α	−0.259	−0.229	−0.193	−0.260	−0.228	−0.240	−0.236	−0.238
rich−club clique, $\phi = 1.0$	13	17	5	21	15	19	16	17
clustering coefficient, $\bar{c}$	0.477	0.286	0.303	0.300	0.483	0.327	0.457	0.312
max degree clique, g	20	20	15	24	21	23	22	21

Table 2 Comparison of PFP and measured datasets in recent years about 2010.

Metrics	ark1210	PFP	bgp1210	PFP	(bgp+ark)1210	PFP	ucla1210	sidewalk
number of nodes, N	29785	29785	36461	36461	36559	36559	36753	31847
number of links, L	78066	88901	100971	108390	128492	108825	115869	143384
maximum degree, K	3405	9106	2939	15985	3896	16051	2983	3709
average degree, $\bar{k}$	5.24	5.97	5.54	5.95	7.03	5.95	6.31	9.00
average shortest path, $\bar{l}$	3.56	2.93	3.92	2.75	3.55	2.81	3.87	3.51
assortative coefficient, α	−0.179	−0.265	−0.210	−0.250	−0.200	−0.238	−0.215	−0.186
rich-club clique, $\phi = 1.0$	15	25	4	26	6	25	4	10
clustering coefficient, $\bar{c}$	0.379	0.377	0.240	0.432	0.462	0.422	0.287	0.42
max degree clique, g	25	28	18	30	26	29	18	19

c_i for node i is defined as the ratio of the actual number of links e_i to the maximum possible number of links among the k_i nearest neighbors of node i, i.e., $c_i = 2e_i/(k_i(k_i - 1))$. The clustering coefficient of a network is the average of local clustering coefficient over all nodes in this network.

- **Max degree clique**. It is the number nodes in the maximum clique that is a full-mesh subgraph created by the following steps. Firstly, the clique contains only one node with the maximum degree. And then, each step chooses a new node in decreasing order of node degrees. If the new node is connected to all the nodes in the clique, it will be added into the clique. Otherwise, discard this node and choose the next one. Repeat this process, and finally get the max degree clique.

5. Results

5.1 The Metrics of AS Graphs

The results are shown in Table 1 and Table 2. *1)* For the assortative coefficient, PFP model has values from −0.265 to −0.229 that are close to the measured datasets in Table 1, but lower than the measured datasets in Table 2, which fluctuate around −0.2. *2)* For the clustering coefficient, we can find that the values of traceroute and traceroute-bgp merged datasets are more than 0.42 except the 0.379 of ark1210, while BGP data keeps a lower value less than 0.303. *3)* The clustering coefficient of PFP model keeps an increasing tendency as network size grows. It is close to the BGP data in the early years around 2002 and the traceroute data in recent years around 2010. *4)* In Table 1, the rich-club cliques of PFP model are close to the traceroute data. In Table 2, the rich-club clique of traceroute data is more than 10, while this metric of BGP data is less than 6. The rich-club cliques of PFP model are larger than the traceroute data, and much larger than the BGP data. This shows that PFP model tends to generate a larger fully-connected rich-club than real AS topology, which is also indicated by the metric *max degree clique*. *5)* PFP model has max degree cliques of size about 20 to 26, similar to the skitter, ark and the bgp+ski/ark merged data. However, the BGP data has smaller values less than 18, and the value of sidewalk dataset is also smaller than PFP, skitter and ark. Table 2 also shows that PFP has a much higher maximum degree, and a notably small average shortest path length. These results present other witnesses that PFP has a more strongly dense core than the Internet AS topology.

5.2 Rich-Club Coefficient

We repeated the experiments in [5] to verify our data and experiment consistency, and got the same result. Figure 1 and the *ski0304* column in Table 1 show that PFP model can accurately fit the data ski0304, as shown in [5] and [7]. We processed other datasets in the same way and obtained the results shown in the following figures. Figure 2

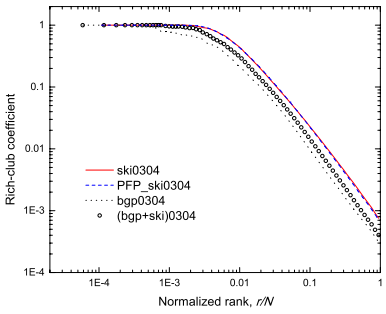

Fig. 1 PFP model can fit the rich-club curve of ski0304 well.

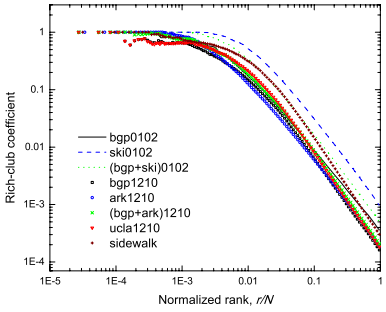

Fig. 2 Comparison of the rich-club coefficient between measured Internet AS graph datasets.

compares the rich-club coefficient of different datasets in Jan. 2002 and Dec. 2010. It can be found that the difference between the curves of ski0102 and bgp0102 is much larger than that between ark1210 and bgp1210. This may be stemmed from the biases in BGP and traceroute data that relies on a smaller number of monitors in the early years about 2002. The curves of bgp1210 and ark1210 are close to each other because both traceroute and BGP measurements cover a wider common area of the Internet with a growing number of monitors in recent years. In the dataset of recent years, the traceroute datasets ark1201 and sidewalk have higher rich-club coefficients than the BGP datasets bgp1210 and ucla1210 when r/N falls into the range about $[0.0001, 0.001]$. This difference may be caused by the missing AS links in BGP data [11]. However, compared with ark1210, we can also find that all the recent datasets, including bgp1210, (bgp+ark)1210, ucla1210 and sidewalk, show raised curves when r/N is around 0.01. This implies a growing number of connections among middle-rich AS nodes. This feature is not notable in ark1210 possibly due to measurement limitations.

Figure 3 shows the comparison between PFP model and the datasets in early years. PFP model has coincident rich-club coefficient with ski0102 and (bgp+ski)0102, but it has visible deviations from BGP data. In Fig. 4, PFP model more or less deviates from the corresponding traceroute and BGP data. We can see that PFP model has a larger rich-club clique than measured data. In PFP model, $\phi(r/N) = 1.0$ when r/N less than 0.001, but measured data doesn't so. However, the datasets bgp1210, (bgp+ark)1210, ucla1210,

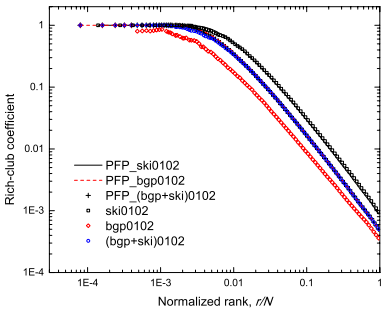

Fig. 3 Comparison of the rich-club coefficient of PFP graphs and the bgp and skitter datasets in 2002.

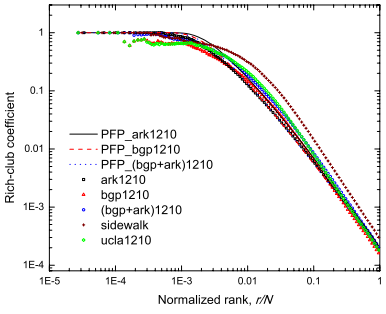

Fig. 4 Comparison of the rich-club coefficient of PFP graphs, the bgp and ark datasets in 2010, and the sidewalk datasets.

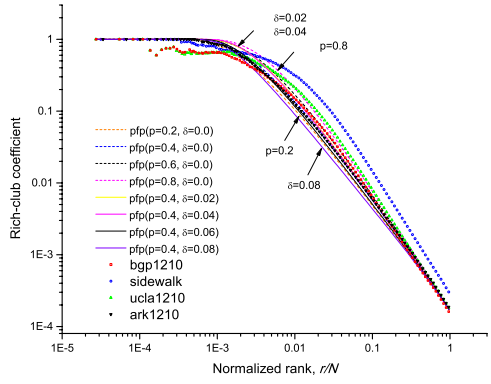

Fig. 5 Comparison of the rich-club coefficient of measured AS graphs, and the PFP graphs with different parameters.

and sidewalk have higher rich-club coefficients than the PFP graphs of the same or similar network size when r/N falls in the range about between 0.01 and 0.1.

To estimate the effect of parameters on PFP model, we generated different PFP graphs of the same size as bgp1210 using different parameters. One group of PFP graphs are generated with fixed $\delta = 0.0$ and varied $p = 0.2, 0.4, 0.6, 0.8$, and another group of PFP graphs are generated with fixed $p = 0.4$ and varied $\delta = 0.02, 0.04, 0.06, 0.08$. Figure 5 shows the rich-club coefficients of PFP graphs of different parameters, and the Internet AS graphs. We can see that PFP graphs cannot accurately fit the recent Internet AS graphs (It seems that ark1210 is very close to the PFP graph of $p = 0.4$ and $\delta = 0.0$, but that PFP graph significantly deviates from ark1210 in other graph metrics, such as maximum degree and assortative coefficient).

6. Conclusion

We used multiple traceroute and BGP datasets to evaluate the rich-club property of the Internet AS-level topology in different views and at different times. It is found that traceroute AS graphs generally yield larger rich-clubs of full connections than that of BGP AS graphs. BGP AS graphs in recent years around 2010 have raised rich-club coefficients $\phi(r/N)$ compared with traceroute AS graphs when r/N falls in the range of 0.01 to 0.1. This implies a growing number of connections among the AS nodes of middle degrees. PFP model shows a more strongly dense core than the measured AS graphs. The rich-club coefficient of PFP model can accurately fit the traceroute AS graphs in the early years around 2002. However, all measured AS graphs in recent years about 2010, including traceroute and BGP datasets, exhibit visible deviations from PFP graphs. These results indicate that topology modeling needs to be improved to capture the variations in the recent and future Internet AS topology.

References

[1] S. Zhou and R.J. Mondragon, "The rich-club phenomenon in the Internet topology," IEEE Commun. Lett., vol.8, no.3, pp.180–182, 2004.
[2] V. Colizza, A. Flammini, M.A. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," Nature Physics, vol.2, no.2, pp.110–115, Feb. 2006.
[3] S. Zhou and R.J. Mondragon, "Structural constraints in complex networks," New Journal of Physics, vol.9, no.173, June 2007.
[4] Z.-Q. Jiang and W.-X. Zhou, "Statistical significance of the rich-club phenomenon in complex networks," New Journal of Physics, vol.10, 043002, April 2008.
[5] S. Zhou. "Understanding the evolution dynamics of Internet topology," Phys. Rev. E, vol.74, 016124, July 2006.
[6] S. Zhou, G.Q. Zhang, and G.Q. Zhang, "Chinese Internet AS-level topology," IET Commun., vol.1, pp.209–214, 2007.
[7] S. Zhou and R.J. Mondragon, "Accurately modeling the internet topology," Phys. Rev. E, vol.70, 066108, Dec. 2004.
[8] Routing Information Service (RIS), http://www.ripe.net/data-tools/stats/ris/routing-information-service, accessed Dec. 2011.
[9] Route Views Project, University of Oregon. http://www.routeviews.org/, accessed Dec. 2011.
[10] UCLA Internet topology collection, http://irl.cs.ucla.edu/topology/, accessed Dec. 2011.
[11] R.V. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, "The (in)completeness of the observed internet AS-level structure," IEEE/ACM Trans. Netw. (TON), vol.18, no.1, pp.109–122, 2010.
[12] AS Links Dataset of the Cooperative Association For Internet Data Analysis (CAIDA), http://www.caida.org/data/active/skitter_aslinks_dataset.xml, accessed Dec. 2011.
[13] P. Mahadevan, D.V. Krioukov, M Fomenkov, X.A. Dimitropoulos, K.C. Claffy, and A. Vahdat, "The internet AS-level topology: Three data sources and one definitive metric," Computer Communication Review (CCR), vol.36, no.1, pp.17–26, Jan. 2006.
[14] Sidewalk Internet AS graph dataset, http://www.aqualab.cs.northwestern.edu/projects/SidewalkEnds.html, accessed Dec. 2011.
[15] Y. Zhang, R.V. Oliveira, Y. Wang, S. Su, B. Zhang, J. Bi, H. Zhang,

and L. Zhang, "A framework to quantify the pitfalls of using traceroute in AS-level topology measurement," IEEE J. Sel. Areas Commun., vol.29, no.1, pp.1822–1836, Oct. 2011.

[16] M.E.J. Newman, "Assortative mixing in networks," Phys. Rev. Lett., vol.89, 208701, 2002.