International Journal of Future Generation Communication and Networking
Vol. 9, No. 1 (2016), pp. 319-328
http://dx.doi.org/10.14257/ijfgcn.2016.9.1.28

A Primary Way of Solving Sampling Bias Problem in Complex Internet Topology

Xu Ye and Shi Wei

School of information science and engineering
Shenyang Ligong University
Shenyang, Liaoning 100159, China
xuy.mail, shiwei}@163.com

Abstract

Measuring complex Internet topology is primary research activities in Internet-related research fields and has become a focus recently. And due to the complexity of Internet topology, further processing is required and is thoroughly studied in this paper. First, Internet topology measurement is performed by CAIDA measurement approach with more than twenty monitors. Then the problem of sampling bias is quantitatively studied. Mathematical models are introduced into both the node bias problem and link bias problem, and final option is gained. Final results are that more than 156 CAIDA monitors are necessary in order to have a full Internet topology measurement result.

Keywords*: Internet topology, Sampling bias problem, Single point measuring, Multi-point measuring*

1. Introduction

Measuring Internet is to accurately capture the quantitative measurement data of the Internet and their activities. Generally, main parameters of network measurement include RTT, path data, bandwidth and delay, congestion, the bottleneck, the target site accessibility, throughput, and bandwidth utilization, packet loss rate, response time of servers and network devices, the largest network traffic, and QoS[1].

Among these network measurements, there is a kind of measurement that is inherent property of a network, such as network topology. And the topology measurement is what this paper studies.

2. Measuring Complex Internet Topology

2.1. Measuring Methods

Static methods based on the BGP route table and the dynamic methods based on the active probing are the main ways to measure the router-level Internet topology [2]. And the static methods are gradually replaced by the dynamic ones due to their lack of the redundant routers measures [2].

The dynamic methods, at present, are mainly divided into three categories [3]:

- Single-monitor-measuring by recording all routers in the route path, such as the Internet Mapping Project (IMP) in Bell Lab.[4], and the Mercator projects [5];
- Active measuring based on the Public Traceroute Server (PTrS), such as the ISP topology measuring project by Boston University [6].

ISSN: 2233-7857 IJFGCN

- Multi-monitor-measuring or measuring-from- multiple-vantage-points by self-developed software engines, such as the CAIDA1 projects [7, 8], and the Active Measuring Project by Harbin Institute of Technology [3].

In the upper three methods, the PTrS (method No.2) is quite limited due to the following reasons [3]. Firstly, PTrS are quite unevenly distributed in Internet and not all ISPs render services of PTrS. Studies in [3] indicated that only one of nine ISPs providing PTrS, so PTrS method is not as reliable as the others. Secondly, it's rather hard to transfer or gain the control of PTrS from the ISPs due to security considerations, which directly resulted in the inefficiency of measuring Internet topology.

The first method is similar to the third one (CAIDA), they are all based on traceroute or the traceroute-like programs [7, 8], but the first method is inferior to the third one since it's totally upon single-monitor-measuring tools. CAIDA, however, could implement multi-monitor-measuring and consequently yield better measuring results [7, 8]. The Active Measuring Project by Harbin Institute of Technology (HIT) also used multi-monitor-measuring tools, but it had fewer monitors in its project than CAIDA has, what's more, the HIT project was mainly focused on the Internet topology in China part [2, 3], on the contrary, CAIDA project measured the world-wide Internet. Therefore, CAIDA method of measurement was used in this paper.

2.2. Measuring Results

The measuring results in this paper are the router-level Internet topology data measured from as many as twenty-one CAIDA monitors. Some of the measuring results are illustrated in the tables as follows.

Table 1. Measuring Results By 21 CAIDA Monitors

No. of monitors	No. of measured IPs	No. of measured routers	No. of measured Links
1	361166	202738	355758
2	412842	278105	546240
3	490940	362262	760571
4	541519	412356	896579
5	600435	470026	1049635
6	652725	521043	1186187
7	697659	562819	1299131
8	750748	612202	1422300
9	801996	659431	1544410
10	860196	713327	1686369
11	906109	755622	1798757
12	945004	789984	1896935
13	991163	831994	2000315
14	1028618	865346	2087219
15	1072059	904805	2185579
16	1111308	939833	2273950
17	1162489	986163	2388690
18	1202696	1022143	2476043
19	1259830	1074065	2602990

1 CAIDA, the Cooperative Association for Internet Data Analysis, is a worldwide research center on Internet-related research fields. CAIDA has more than thirty monitor nodes which are distributed throughout the whole world, measuring and monitoring the variations of Internet. Three of the monitors are located in Asia.

20	1295194	1105843	2695032
21	1307423	1116474	2717358

It's clearly seen that the measured results are getting more with the increment of monitors. For a definitely same Internet topology currently in front of us, why there are different measuring results coming out for us? Are the measuring results good enough for our experiments? To answer these questions, further processing such as Sampling Bias [3, 6] are to be introduced to these topology samples.

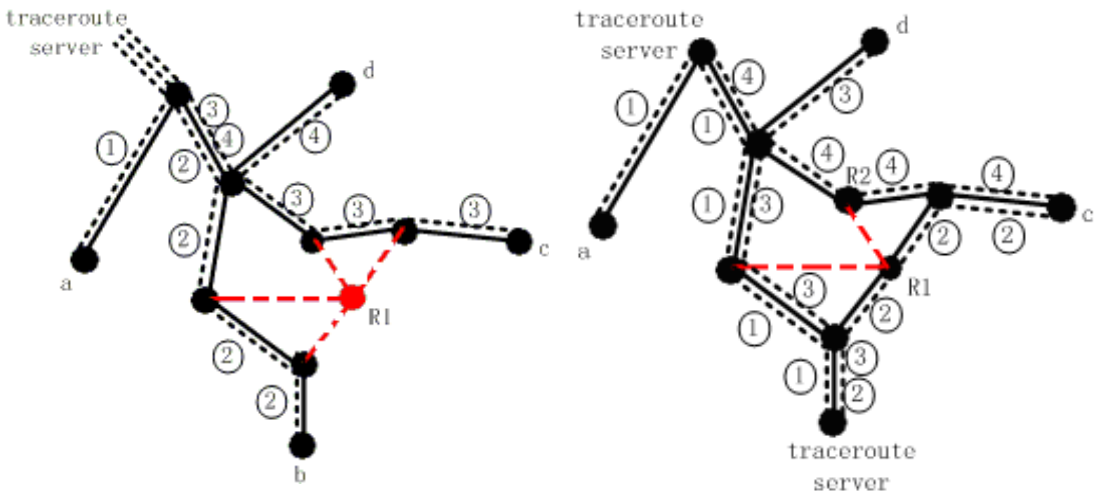
3. Processing Complex Internet Topology

3.1. Problems of Sampling Bias

Some recent researches [3, 6] found that the measuring results were usually different from real network topology and tended to show stronger power-law (frequency-degree power-law) properties than what the real network actually has when only one monitor or less monitors was used during the active measuring by the traceroute-like tools. And this is called sampling bias [6].

Sampling bias occurs when a handful of monitors in are used in detections of a target system with huge nodes and links, and results in lack of monitored nodes and links. Then, we'll study how much noise the sampling bias would add to the monitored Internet topology, and the way to solve it.

Some recent researches[6] found that the measuring results were usually different from real network topology and tended to show stronger power-law (frequency-degree power-law) relations when only one monitor or just a few monitors was used during the active measuring. For instance, one measuring monitor paradigm is illustrated in Figure 1(a).



(a) Measuring a target network with four nodes (a, b, c and d) from one monitor with traceroute-like tools. The measure covers four path indicated by (1-4). The dotted links and R1 are the missing routers and links for sampling bias.

(b) Measuring the three leaf nodes (a, c and d) from two traceroute monitors. The covered path are indicated by (1-4). The dotted links are the missing routers and links.

Figure 1. Illustrations of Measuring a Network From Different Monitors

From Figure 1(a), Router R1 and four links (the dotted links) are missed out. And difference between the measuring results from the real network is known as sampling bias [6]. Sampling bias is directly associated with the number of measuring monitors [6, 9]. To prove this, let's go on experiments illustrated in Figure 1(b), which has two monitors.

From Figure 1(b), Router R1 and two links missed in Figure 1(a) were successfully found. But there are still two dotted links missed due to sampling bias. Though it's still hard to find perfect approaches solving the sampling bias problems at present [6, 9], we still found an easy and effective way from the last two figures. To solve, in some extent, the problem of sampling bias, it is helpful to use more monitors in measuring target network. And this is also the way we used in this paper.

3.2. Quantitative Analysis of Sampling Bias

1) CAIDA measurement results

For better analysis of the effect of sampling bias on Internet measuring results, we studied all twenty-one measurement data. And the relationship among the monitoring number, monitored nodes and monitored links are illustrated in Figure 2. And the relationship between the increment of reached links, nodes and that of monitoring numbers are illustrated in Figure 3.

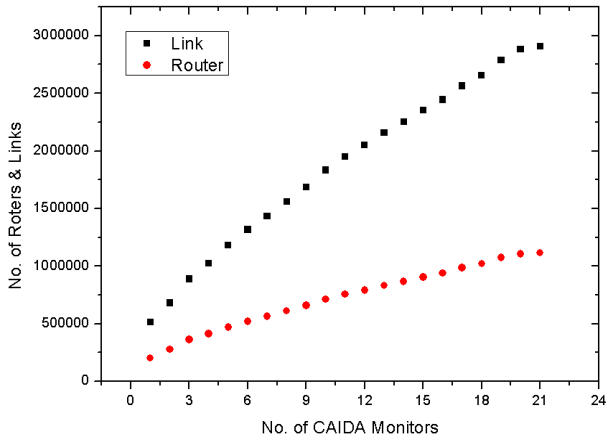


Figure 2. Growth of the Reached Routers and Links

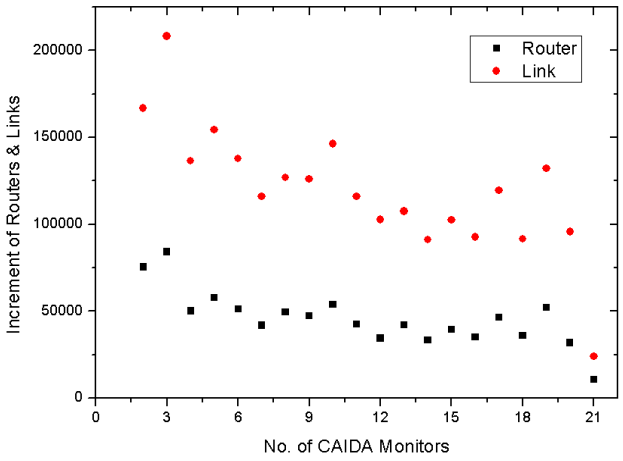


Figure 3. Growth of Increment of Reached Routers and Links

From Figure 2, it's clear that with growth of monitoring numbers, both reached nodes and links are found to increase significantly. What's more, the increment of links is obviously stronger than that of nodes. Probable reasons are that each new found node (router) will always lead to more new found links (routed connections). Figure 1 gives only an overall description of the relationships. And Figure 2 gives details.

From Figure 3, first the nodes, though the increment varies much, the decreasing trend is clearly found during its growth. The decaying increment indicates that the most significant newly-found-node increase comes when CAIDA monitor number is rather small, and with more newly monitors put into measurement activities, the measurement result approach the real target of network – Internet, the newly found nodes keep decreasing. And this is quite consistent with our common sense. For the link increment, the growth situation is exactly the same.

2) Mathematical analyses of node increment

Assume the number of overall missed nodes in measuring Internet is M , and the newly-found-node increment is r_n when adding n new monitors. Then we have

$$\lim_{n->\infty} (\sum_{i=2}^{n} r_i) = M \qquad \textbf{(1)}$$

Considering that no matter how many new monitors added in measurement, there are always ε nodes missed. Then

$$\lim_{n->\infty} (\sum_{i=2}^{n} r_i) + \varepsilon = M \qquad \textbf{(2)}$$

So, if we try to acquire the mathematical model of r_n , there will be a way to find M .

[Proposition 1] When adding new monitors in Internet measurement, mathematical model of the node increment r is a generalized decreasing function.

Proof:

Use of reduction to absurdity. Assume r is generalized increasing function, then $r_i < r_{i+1}$ (i>=2), so we have $\lim_{n->\infty} (\sum_{i=2}^{n} r_i) > \left(\lim_{n->\infty} (\sum_{i=2}^{n} r_2) = \lim_{n->\infty} (nr_2) \right)$, where r_2 is node increment when adding one monitor (in total is two).

According to Figure 1 and objective common sense, r2>1, then $\lim_{n->\infty} (nr_2) = \infty$, Combine the upper two equations with equation 2, we have $\left(M = \lim_{n->\infty} (\sum_{i=2}^{n} r_i) + \varepsilon \right) > (\infty + \varepsilon = \infty)$.

Then $M > \infty$ is gained. M , however, denoted the missed routers in Internet, is always a limited quantity, $M < \infty$. So absurdity is found here. Thus, mathematical model of the node increment r is a generalized decreasing function.

End.

From Figure 1, we can find that r is decreasing with the course of oscillation. If so, r_{i+1} is not definitely less than r_i . Since r is proved to be decreasing in the long term, there must be sufficiently large positive integers W and l to make certainty of the followed equation.

$$\frac{\sum_{i=k}^{k+w} r_i}{W} > \frac{\sum_{i=k+l}^{k+l+w} r_i}{W} \quad (W>0,\ l>0) \qquad \textbf{(3)}$$

And with the same approach as proof of Proposition 1, same results of decreasing function of r will also be gained.

3.3. Fitting of Node Increment

For better results, ten kinds of fitting models are used in fitting node increment r. And they are: Liner model, Quadratic model, Compound model, Growth model, Logarithmic model, Cubic model, S curve, Index equation model, inverse curve model, and Power model.

The fitting results are illustrated in Figure 4. From Figure 4, all ten models could fit the observed data well. Further analyses are required and the results are shown in Table 1 below.

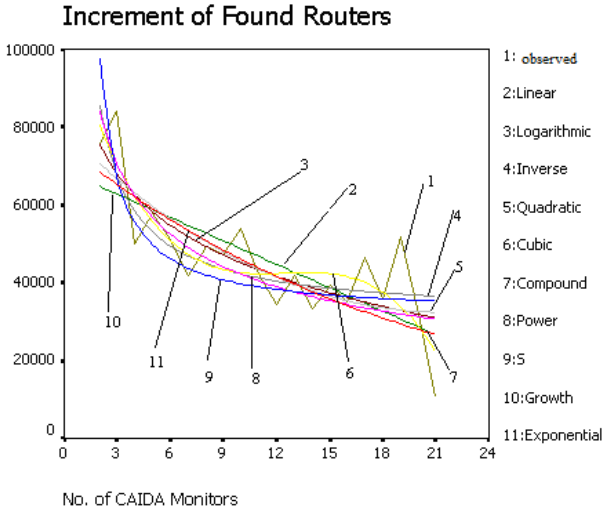


Figure 4. Fitting Result of Node Increment with Growth of the Number of Monitors

Table 2. Fitting Results of Node Increment with Growth of the Number of Monitors

Mth	Rsq	d.f.	F	Sigf	b0	b1	b2	b3
LIN	.574	18	24.29	.000	68770.8	-2007.3		
LOG	.655	18	34.13	.000	88782.0	-18993		
INV	.625	18	29.95	.000	31365.2	108277		
QUA	.613	17	13.46	.000	78796.7	-4336.6	101.272	
CUB	.727	16	14.19	.000	107763	-16032	1298.21	-34.694
COM	.506	18	18.41	.000	75717.5	9514		
POW	.478	18	16.45	.001	113045	-.4292		
S	.385	18	11.28	.003	10.3641	2.2501		
GR	.506	18	18.41	.000	11.2348	-.0498		

O P	EXP	.506	18	18.41	.000	75717.5	-.0498

From Table 2, it's found that all ten models have statistical significances. Probability of variance analyses F are rather small, approximate to 0, the largest is only 0.003. For R2, R2 (CUB) =0.727 is the largest. Then the cubic model is found to be the best for fitting r . And its parameters are b0=107763, b1=016032, b2=1298.21, b4=-34.694.

It seems that the cubic model is the final option. However, if substitute 30 (the number of monitors) into cubic parameter, the node increment is a large negative integer, which is totally wrong.

So the option of fitting model relies not only on the fitting accuracy, but on the adaptability.

With this rule, the Logarithmic model is the final option. And the model is:

$$y = 88782.0 - 18993 \times \ln(x)$$

(4)

According to Eq. 4, we set y=0, and result is x=107. Meaning that when the number of CAIDA monitors reaches 107 or above, there will be no nodes missed.

3.4. Fitting of Link Increment

Similarly, fitting analysis of the link increment is performed and the results are illustrated in Table 3 and Figure 5.

Table 3. Fitting Results of Link Increment with Growth of the Number of Monitors

Mth	Rsq	d.f.	F	Sigf	b0	b1	b2	b3
LIN	.599	18	26.93	.000	175044	-4813.2		
LOG	.611	18	28.28	.000	217423	-43072		
INV	.504	18	18.26	.000	89507.3	228209		
QUA	.607	17	13.14	.000	185606	-7267.0	106.687	
CUB	.670	16	10.85	.000	236315	-27741	2202.10	-60.737
COM	.469	18	15.87	.001	197174	.9522		
POW	.401	18	12.04	.003	279240	-.4019		
S	.283	18	7.11	.016	11.3673	1.9714		
GRO	.469	18	15.87	.001	12.1918	-.0490		
EXP	.469	18	15.87	.001	197174	-.0490		

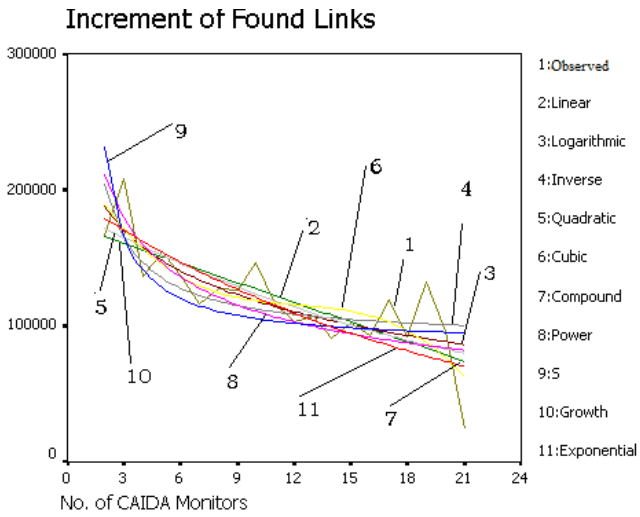


Figure 5. Fitting Result of Link Increment with Growth of the Number of Monitors

Similarly, the option of the link increment model is Logarithmic model with parameters as follows:

$$y = 217423.0 - 43072 \times \ln(x) \quad \textbf{(5)}$$

Using Eq. 5, the solution of x is 156 when y=0. Meaning that when the number of CAIDA monitors reaches 156 or above, there will be no nodes missed.

According to Eq. 4 and Eq. 5, we get x=max (107, 156) =156. Which means that if no node and link missed in Internet measurement, at least 156 CAIDA monitors is required.

For all measured nodes and links, wet get 156 monitors is our final result.

4. Conclusions

In this paper, Internet topology measurement is performed by CAIDA measurement approach with more than twenty monitors. Then, due to the complexity of Internet topology measurement result, further processing is required - the problem of sampling bias is quantitatively studied.

Finally, mathematical models are introduced into both the node bias problem and link bias problem, and Logarithmic model is found to be a suitable fitting model. With the mode, it is solved that more than 156 CAIDA monitors are necessary in order to have a full Internet topology measurement result.

Acknowledgement

This work is financially supported by the National Natural Science Foundation of China (No. 61373159), the Shenyang Natural Science Foundation (F13-316-1-22) and the Open foundation of Key lab of Information Networking and Confrontation of Shenyang Ligong University (No. 4771004kfs18).

References

[1] H.L. Zhang, B.X. Fhang and M.Z. Hu, "A Survey on Internet Measurement and Analysis [J]", Journal of Software, vol. 14, no. 1, **(2003)**, pp. 110-116.
[2] B. Huffaker, D. Plummer and D. Moore, "Topology discovery by active probing [EB/OL]", http://www.caida.org/outreach/papers/ 2002/SkitterOverview/, **(2002)**.
[3] J. Yu, F. Binxing and H. Mingzeng, "Mapping Router-level Internet Topology from Multiple Vantage Points [J]", Telecommunications Science, no. 9, **(2004)**, pp. 12-17.

[4] B. Cheswick, H. Burch and S. Branigan, "Mapping and visualizing the Internet [C]", In: Proc of the 2000 USENIX Ann Technical Conf, San Diego, California, USA, **(2000)**.
[5] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet map discovery [C]", In:Proc of IEEE INFOCOM, **(2000)**.
[6] N. Spring, R. Mahajanand D. Wetherall, "Measuring ISP topologies with rocketfuel [J]", ACM SIGCOMM Computer Communication Review, vol. 32, no. 4, **(2002)**, pp. 133-145.
[7] Skitter, CAIDA, http://www.caida.org/tools/measurement/skitter/
[8] "Mapnet: Macroscopic Internet Visualization and Measurement", CAIDA. http://www.caida.org/tools/visualization/mapnet/
[9] A. Lakhina. J.W. Byers, M. Crovella and P. Xie, "Sampling biases in IP topology measurements [C]", In: Proc. of the IEEE INFOCOM 2003, San Francisco: IEEE, vol. 1, **(2003)**, pp. 332-341.

