Taking Too Many Destinations Can Be Bad for Traceroute Sampling

Zhongliang Qiao, Mingming Chen, Ke Xu*
State Key Lab. of Software Development Environment, Beihang University, Beijing 100191, P.R.CHINA
{qiaozl, chenmingming, kexu}@nlsde.buaa.edu.cn

Abstract—**Considerable effort has been spent on collecting all the information of routers by using traceroute-like probes in the router-level topology measurements. This method has been argued to introduce uncontrolled sampling biases on statistical properties of the sample graph and heavy load to the network being measured. In order to improve the quality of the maps induced by the method, researchers are starting to investigate the deployment of large-scale distributed systems. But the lack of sources, the additional load introduced to the network and the potential uncontrolled scale in the IPv6 network cast a shadow over this direction. In this paper, we study *traceroute sampling* to represent the topology. Instead of finding a general strategy that would match all the graph properties, we focus on testing the impact of the proportion of destinations and sources on a single or several properties of the graph. We argue that, in order to obtain a more accurate sample graph, the general method of taking a small set of sources to perform traceroute-like probes to a large set of destinations is very unreasonable. Our results obtained from simulated experiments show that as the proportion of destinations and sources increases, the resulting properties of the sample graph can differ sharply from the underlying graph. The results also show that there is no single perfect proportion answer to meet all the graph properties but the small often perform better than the large overall. When we do the same measurement on several real-world networks, we find strong evidence for sampling bias because of taking so many destinations.**

Keywords-component: network topology; graph sampling; traceroute

I. Introduction

SINCE the Internet was constructed and managed in a fully distributed manner, it is hard to obtain a high accurate map of the Internet which plays a vital role in key problems, e.g., network robustness [7][8], simulation of protocols and applications [9]. Instead of accurate map, researchers currently rely on various probing methods to assemble an approximate map of the topology in the router-level. The main approach of such strategy is to use traceroute. In brief, a limited set of sources perform traceroute-like probes to a large set of destinations, then merge these traces into one graph. This technique has been applied to many famous topology discovery systems [1][4][5][6].

Recently, several works have focused on evaluating the accuracy of the obtained maps of the Internet [2][3][10][11]. All these studies recognized the fact that currently available maps of the Internet induced by traceroute are very incomplete, and an important bias may have been induced by the exploration process. Researchers also raised the problem of this approach that unless be carefully controlled, the discovery systems have the potential to introduce a heavy load to the part of the network being measured [15]. They also have the potential to raise alarms, as their traffic can easily resemble a distributed denial-of-service (DDoS) attack. Both of the drawbacks are difficult to overcome.

Currently, in order to improve the quality of these maps induced by traceroute, researchers are starting to investigate the deployment of large-scale distributed systems [12][13][14]. We have also provided a logic distance-based method for large-scale deployment of traceroute-probing sources [23]. Obviously, there are two obstacles for this strategy. One is the difficulty in deploying a large number of traceroute apparatuses; the other is the even heavier load brought to the network. Further, the situation would be far worse if we take the measurements of the next generation of network (IPv6) into consideration. The IPv6 network will be huge which may make the scale of the distributed measurement systems and the load introduced to the network out of our imagination. So we can foresee that the direction of collecting all the information of routers using traceroute to do large-scale census is faced with great difficulty.

Different from previous studies which are sparing no effort on collecting as much information of the routers as possible, we explored the possibility of traceroute sampling in observing the network topology. The sampling problem is defined as: Given a graph *G=(V,E)*, then randomly choose *k* sources and *m* destinations from *V*, let *(k,m)-traceroute* denote the traceroute-like probes exploration from *k* sources to *m* destinations. The task is to use *(k,m)-traceroute* to create a small sample graph *G'* that will be similar (in properties) to *G*. We study the impact of the proportion (*m/k*) of destinations and sources on the accuracy of *G'* in order to test the impact of *m/k* in estimating *G*. To validate the similarity of the sample graph, we reexamine a set of properties of *G'* and compare them with *G*. Of course, the placement of sources and destinations also influence the accuracy of *G'*. But note that the aim of our work is not to investigate strategies of using *(k,m)-traceroute* to sample a perfect *G'*, we are more interested in the ability of a simple *(k,m)-traceroute* representing the network topology.

As mentioned above, several works have evaluated the accuracy of the network induced by *(k,m)-traceroute*, but the authors hope to collect as much information of the topology as possible to study the bias induced on the degree distribution. They do not analyze the correlation between the proportion of destinations and sources and the accuracy of the degree

*Corresponding author. Tel.:+8610 8231 5704; E-mail:kexu@nlsde.buaa.edu.cn distribution, not to mention the other properties of the graph.

The main results of this work are the following:

We analyze the correlation between the proportion of destinations and sources and the accuracy of *(k,m)-traceroute* sample in perspective of several properties of the graph. Our results show that the proportion of destinations and sources has a great impact on the accuracy of the sample graph. The general method of taking plentiful destinations while a small number of sources can cause the sample graph differ sharply from the underlying graph in perspective of several properties of graph. The results also suggest that there seems to be no single perfect *m/k* answer to meet all the properties of graph and the small *m/k* often perform better than the large overall.

The rest of this paper is organized as follows. In Section II, we present the models we use and define the properties and methods in evaluating the sample graph. Then we present and analyze the results of our simulations on various models and statistical properties in Section III. Section IV devotes to check the performance of different *m/k* on several real-world networks. Finally, we make conclusion of our experiments and discuss some of future work in Section V.

II. Preliminaries

In study of traceroute sampling, we are given a large target graph that represents the network. The task is to create a small sample graph that will be the most similar to the original graph. The tool we use is traceroute and the process for sampling is *(k,m)-traceroute* exploration. The similarity of the sample graph will be evaluated according to several criteria using a common pattern matching technique.

A. The target graphs for sampling

A network topology can be naturally represented as a graph. In our study, the graph does not need to be weighted nor directed. In modeling such a graph, three characteristics have been considered most: low average distance, scale-free and high clustering [3]. According to these characteristics, a lot of models to generate the graph have been proposed.

The basic model of the network is the Erdős-Rényi (ER) [16] model. In such a model, every two nodes are connected to each other with a probability *p*. Though the average distance is low, the degree distribution of the model is Poisson. BA [17] model makes a further step. It is based on the preferential attachment. The proportion p_k of nodes with degree *k* follows a power law with an exponent γ: $p_k \sim k^{-r}$. The average distance of such a graph is logarithmic in the number of nodes. However, the clustering of ER and BA models is very low which lost another important characteristic of the network. This was improved by the extended BA model (EBA) [18]. Because the preferential attachment is hidden in this model and the node added into the graph during the growth process forms a triangle at each step, the clustering of the graph is much higher. The degree distribution of the EBA model also follows the power law.

The three models cover all the three important characteristics of the network. So we believe it has been enough to carry out our research. The graphs generated based on these models in our research will be introduced in the next section.

B. Modeling traceroute and (k,m)-traceroute exploration

Additional to the graph models, we need the routing model as viewed by traceroute and the exploration process model of *(k,m)-traceroute*. A route in the network given by traceroute is a path in the corresponding graph. Because modeling the complex IP routing is beyond the scope of this paper, in our study, we also take the classical assumption [2][3] that a route obtained by traceroute is nothing but a shortest path between the source and the destination. In this condition, for the exploration process of *(k,m)-traceroute*, work[19] proposed three mechanisms: USP (Unique Shortest Path), RSP (Random Shortest Path) and ASP (All Shortest Path). Indeed, for a long-time exploration, *(k,m)-traceroute* may contain a mixture of the three mechanisms. However, because traceroute sampling only takes one-time running of *(k,m)-traceroute* to make a snapshot of the network, we choose USP policy which is the closest to one-time running of *(k,m)-traceroute* as our exploration model.

C. Evaluation techniques

Criteria for a graph

As mentioned above, scale-free, low average length and high clustering are the three most important characteristics of the network discovered. In this part, we present three properties concerned with these characteristics. Given a graph, we measure all the following three properties. Essentially we treat them as distributions to allow for proper scaling:

Degree distribution (*dd*): For every degree *d*, we count the number of nodes with degree *d*. Degree distribution provides a detailed view of the structure of a network and is one of the most frequently used topology characterization metrics [20].

Clustering distribution (*cd*): Clustering provides a measure on how close a node's neighbors are to forming a clique. The larger the local clustering of a node, the more interconnected are its neighbors [20]. The distribution of the clustering coefficient C_d is defined as follows. A node *v* with *k* neighbors could have edges up to *k(k-1)/2* between them. Let C_v denote the fraction of these edges that actually exist. Then C_d is defined as the average C_v over all nodes *v* of degree *d*.

Shortest path length distribution (*spld*): The shortest-path length distribution is a strong indicator of network performance as it shows the reachability of nodes within each other, and for viruses and worms spreading over portions of the network [20]. For every length *l*, we count the number of the shortest paths over the graph with length *l*.

Formal test for criteria

In order to compare the sample and original graph patterns, we use the Kolmogorov-Smirnov *D*-statistic [21] which is usually applied as a part of Komogorov-Smirnov test to reject the null hypothesis. Let *x* denote the random variable over the range and $F'(x)$ and $F(x)$ are two empirical cumulative distribution functions of the data. The *D*-statistic is defined as $D = max_x\,\{|F'(x) - F(x)|\}$. From the definition, we can see that the *D*-statistic represents the

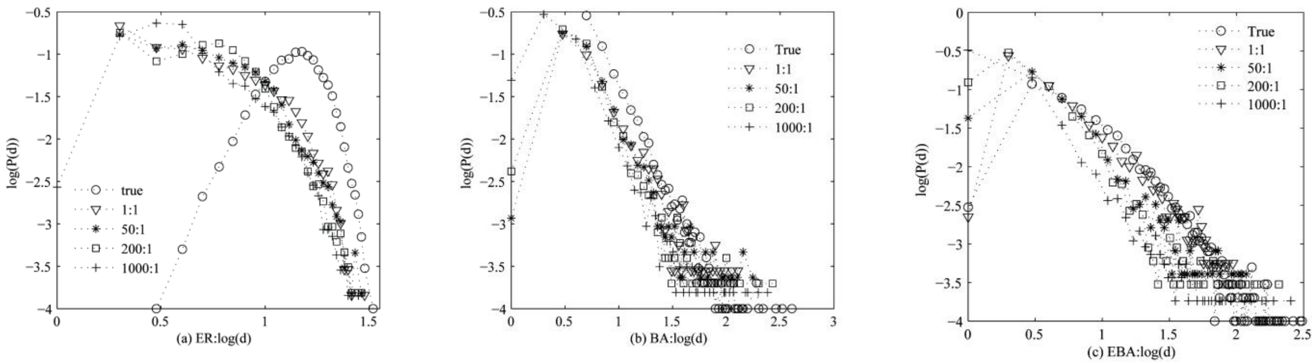


Fig.1. Degree distribution (log-log) of the sample graphs as *m/k* varies. *True* represents the underlying graph, *d* denotes the degree, *P(d)* denotes the frequency of nodes with degree *d*

maximum vertical deviation between the two curves. Here we simply use it to measure the difference between the two distributions. Note that the *D*-statistic does not address the issue of the scaling but rather compresses the shape of the distribution.

Evaluation procedure: Given a sample graph *G'* and a target graph *G*, we first measure the three distributions on both *G'* and *G*, then compare them using the *D*-statistic.

III. Experimental Evaluation

In the following section we will present the results of the experiments on several simulated graphs. We first introduce the generated graphs used to carry out our research. Then, for the experiments, we first present the illustrative examples of the behavior as *m/k* varies and then evaluate the behavior against all the three properties. To test the scalability of each *m/k*, we also explore how the quality of the sample graph changes with the sample size.

A. Graphs description

We generate three graphs based on the models: BA, EBA and ER. The parameters are set to generate the graphs that are similar to the Internet. Their typical values [7] are listed in table I. The graphs are generated using the tool network manipulator (*nem*) [22], version 0.9.6. The graph generated from the tool is in the format of link file of the graph. As mentioned in the previous section, for our purpose, the graph does not need to be weighed nor directed. So we simply assign every link with weight 1. Then we use the link file as the input of our experiments. Every time we run *(k,m)-traceroute* exploration, first we randomly select *k* sources and *m* destinations from the graph. Second, we perform *Dijkstra* algorithm to calculate the shortest paths from each source to all destinations. After these procedures, we obtain the raw trace data. Analysis is performed on these raw trace data sets.

B. Matching the graph patterns

TABLE.I.
Parameters for generating the graphs

	BA		EBA					ER	
parameter	n	m	n	m0	m	p	q	N	P
value	10^4	5	10^4	3	2	0.6	0.1	10^4	1.5^{-3}

In this part, we present some illustrative examples for the behaviors of different *m/k*. We set 10 values of *m/k* ranging from 1 to 1,000, and set 10 different sample sizes (the number of shortest paths) ranging from 5,000 to 20,000($m \times k$). For each *m/k* on each graph, we run *(k,m)-traceroute* exploration on all the given sample sizes.

Figure 1 plots the *dd* on log-log scale of the sample graphs. Each sample is obtained by merging 20,000 shortest paths (the results obtained on the other given sample sizes are similar). Because the lack of space, we do not plot the results of all the 10 proportions (*m/k*). We notice the same behavior on three graphs that, when *m/k* increases from 1 to 1000, the nodes with high degree would be gradually underestimated and finally nodes with low degree like 1 and 2 will be overestimated. For example, figure 1(a) shows that the subgraphs sampled from the ER graph are more like power-law graphs. One qualitatively behavior of *m/k* on the BA and EBA graphs is also found. Since it is difficult to distinguish from figure 1(b)(c), we further draw their trend lines of the graphs in table II. For each trend line, the slope denotes the estimation of the exponent γ and R^2 denotes the correlation coefficient. Table II shows that on the BA and EBA graphs, when *m/k* is small, the estimation of the exponent would be smaller than the truth; while as *m/k* increases, the exponent would gradually increase and finally overestimate the truth.

The result in table II tells that in estimating the scale-free property of the BA and EBA graphs using *(k,m)-traceroute* sampling, the number of sources and destinations should be properly set to a specific value which should be neither too large nor too small .

C. Comparing graph patterns using the D-statistic

Next, we evaluate *m/k* against all the three properties on each

TABLE.II.
Trend lines of scale-free graphs as *m/k* varies.

	BA		EBA	
m/k	Trend line	R^2	Trend line	R^2
1	y = -2.074x + 0.283	0.983	y = -1.728x + 0.049	0.959
50	y = -2.456x +0.583	0.959	y = -2.110x + 0.311	0.938
200	y = -2.545x + 0.603	0.951	y = -2.400x + 0.487	0.939
1000	y = -2.690x + 0.603	0.963	y = -2.462x + 0.194	0.956
True	y = -2.632x + 1.484	0.973	y = -2.026x +0.560	0.973

TABLE.III.
Evaluate criteria using *D*-statistic

m/k	ER				BA				EBA			
	dd	cd	spld	AVG	dd	cd	spld	AVG	dd	cd	spld	AVG
1	0.699	**1.90E-4**	**0.429**	0.376	0.683	1.33E-4	**0.109**	0.264	**0.136**	1.20E-3	0.108	0.082
10	**0.643**	2.43E-3	0.463	**0.369**	0.669	6.82E-5	0.126	0.265	0.148	7.00E-4	**0.096**	**0.081**
20	0.649	5.45E-4	0.483	0.378	0.590	4.58E-5	0.131	0.240	0.158	6.00E-4	0.134	0.098
50	0.683	2.43E-3	0.492	0.392	0.362	9.26E-5	0.133	0.165	0.254	8.00E-4	0.151	0.135
100	0.685	2.43E-3	0.508	0.398	**0.299**	2.72E-5	0.132	**0.143**	0.238	4.00E-4	0.112	0.116
200	0.699	5.44E-4	0.514	0.404	0.309	2.50E-5	0.145	0.151	0.287	1.00E-4	0.154	0.147
400	0.701	6.81E-3	0.530	0.412	0.323	4.82E-5	0.171	0.164	0.365	4.00E-4	0.218	0.194
500	0.700	5.45E-4	0.540	0.413	0.327	4.20E-5	0.176	0.167	0.382	2.00E-4	0.217	0.199
800	0.696	4.13E-3	0543	0.414	0.341	**2.05E-5**	0.208	0.183	0.425	100E-4	0.231	0.222
1000	0.703	1.22E-3	0.554	0.419	0.350	2.66E-5	0.221	0.190	0.445	**0.00E-4**	0.342	0.262

graph. We present the result in table III. Each entry in the table is obtained by averaging the *D*-statistic over 10 different sample sizes.

For each column, we bold the best scoring *m/k*. Notice that on the ER and EBA graphs, almost the smallest *m/k* fits the best on the *dd* and the sampling bias would increase as *m/k* increases; While on the BA graph, middle *m/k* fit better than the smallest and the largest and the sampling bias in this column firstly decreases and then increases. In fitting the *cd*, the large *m/k* perform better on the scale-free (BA and EBA) graphs; while on the ER graph, the smallest *m/k* performs best. There is not much difference among different *m/k* in fitting the *spld* on the three graphs. That is the smaller *m/k*, the better fitting of the original graph.

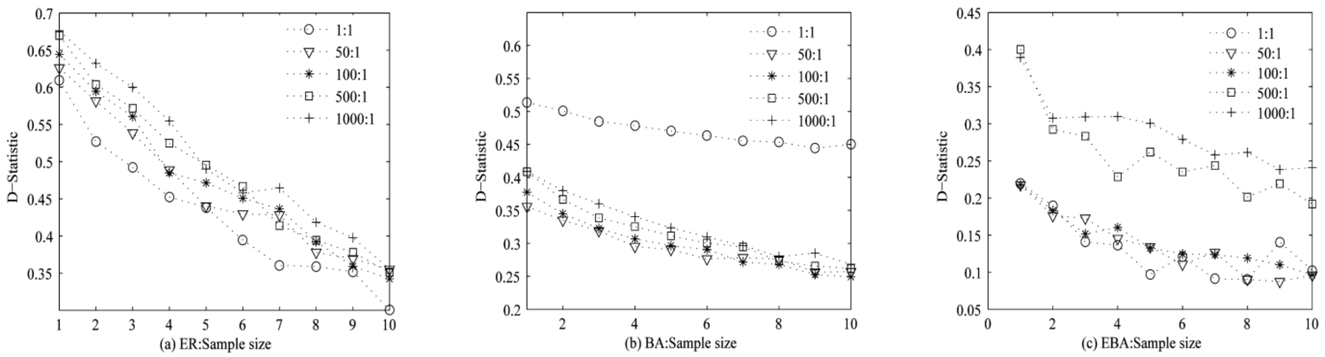
On the BA graph, the *D*-statistic of the *cd* and *spld* change little as *m/k* varies compared to the *D*-statistic of *dd*. So, in sampling the BA graph, it suggests that the *m/k* should be set to an appropriate value in better fitting the *dd* so that obtain the best performance overall. In sampling the EBA and ER graphs, the *m/k* should be set as small as possible. The gain in minimizing *m/k* is especially significant in sampling the EBA graph in perspective of the *dd* and *spld*.

The result in table III tells that, in perspective of the three properties of graph, the *m/k* should be set to a specific value or even as small as possible. Together with the results we obtained in the previous part (part B), the conclusion is sensitive because in practice, in order to obtain a more complete map of the network, the measurement tools simply use a large number of destinations while a small number of sources which makes *m/k* huge. For example, currently, the *CAIDA Skitter* project [1] takes thirty geographically distributed probing sources performing ICMP-based traceroute to approximate 971,000 destinations. The number of destinations is still in growth. If we assume that the network is more like the EBA graph, the exponent may be significantly overestimated by the sample graph, and the *dd* and *spld* may also differ sharply from the underlying graph.

D.The performance of m/k as sample size varies

An appropriate *m/k* should not only perform well on a fixed sample size, but also have a good scalability. Finally, we will explore how the quality of the sample graph changes with the sample size. Figure 2 shows the performance of different proportions of destinations and sources. We plot the *D*-statistic as a function of the sample size. The *D*-statistic is the average of the *D*-statistic of *dd*, *cd* and *spld*. Because the lack of space, we still do not plot the results of all 10 proportions (*m/k*).

Figure 2(a) shows that on ER graph, different *m/k* have an approximate scalability as the sample size varies. Notice that, though the quality of the sample graph is poor when the sample size is small, it will be greatly improved as the sample size increases. Figure 2(b) shows that, on the BA graph, *(k,m)-traceroute* exploration performs worst with the smallest *m/k* and even up to 10-time sample size of the initial, they perform about the same. Also notice that the quality of the sample graph improves as *m/k* increases from the smallest to a threshold (from 1 to 50); while it will slowly decreases when it is over the threshold (from 50 to 1000). It implies that we should properly select the value of *m/k* in sampling the BA graph to achieve the best accuracy overall. On the contrary, figure 2(c) shows that, on the EBA graph, *(k,m)-traceroute* exploration with the smallest *m/k*



(a) ER:Sample size
(b) BA:Sample size
(c) EBA:Sample size

Fig.2. The average *D*-statistic of the *D*-statistic of *dd*, *cd*, *spld* as the sample size($x \times 10^4$) varies.

often performs the best overall even with a small sample size. At the same time, it is not essential that *m/k* is strictly limited to the smallest. Because the performances of the sample graphs with *m/k* range from 1 (even larger) to 100 are approximate to the best *m/k*. But once *m/k* becomes larger, if you want to sample a good graph, you have to select a large sample size.

We have shown that different *m/k* have significant impact on the accuracy of the sample graphs. So in sampling the network using *(k,m)-traceroute*, we would like to ask: what is the best proportion of destinations and sources? There seems to be no single perfect *m/k* answer. For example, if we treat the network as the EBA graph, in estimating the exponent of the Internet, we should set the *m/k* ranging from 1 to 50(see table II) as observed in our experiments; however, in estimating the three properties, different proportions have different performance and how they influence the other properties that we did not include in this study is still unknown. Furthermore, these results suggest that the method of simply increasing the number of destinations while taking a small number of sources could make the observation of the exponent and several properties of the sample graph differ sharply from the underlying graph while the small *m/k* often perform better than the large overall; this effect is especially significant when the sample size is small. In the next section, we will validate this conclusion on two real-world networks.

IV. EXPERIMENTS ON REAL-WORLD NETWORK

In section III, we have shown the simulated results on several typical models by using the simple traceroute-exploration model. In this section, we will further explore the correlation on some real-world networks by making exactly the same measurements.

First, we should select the data sets. There are many famous data sets [1][6][13] obtained by traceroute exploration. For our aim, these data sets must share a common relative abundance of sources and destinations, and at the same time, the maps generated from the data sets must have all the three properties we have mentioned: low average distance, scale-free and high clustering. For this reason, we select the two data sets in our experiments: *Dolphin*2[24] and *iPlane*[25]. Among the sources of the two data sets, many are not available or could reach only a small number of destinations. We use a greedy method to select available sources and destinations as many as possible. The method is: select one source promising that it shares the most reachable destinations with other sources previously selected. Based on this method, the summary of the two data sets for our experiments are listed in table IV. Secondly, in simulating the (*k*,*m*)-*traceroute* exploration, we should take the same sources and destinations as in the real exploration, and we should take the real routes rather than the shortest paths because only this way that is fully respect to the real scenario of the Internet during (*k*,*m*)-*traceroute* exploration. Note that Dolphin2 takes only one single probe engine in conjunction with source routing facilities as source IPs and therefore the mechanism of (*k*,*m*)-*traceroute* exploration cannot be compared with the one deploying traceroute apparatus. However, this method has been applied to some famous projects like *Mercator*[4] and *Atlas*[5]. So it is meaningful to study this kind of (*k*,*m*)-*traceroute* .

TABLE.IV.
Summary of the two data sets

	Dolphin2	iPlane
Source number	362	144
Destination number	1738	12,000
Obtained date	4/12/2009	4/15/2009
Network type	IPv6	IPv4

TABLE.V.
Trend lines of real-network graphs as *m*/*k* varies.

	Dolphin2		iPlane	
m/k	Trend line	R^2	Trend line	R^2
2	**y=-2.26x+3.68**	0.92	**y=-2.92x+4.92**	0.92
8	y=-2.37x+3.79	0.93	y=-3.05x+5.04	0.94
26	y=-2.41x+3.72	0.96	y=-3.09x+5.10	0.96
106	y=-2.54x+3.88	0.96	y=-3.23x+5.24	0.95
425	y=-2.67x+4.23	0.98	y=-3.34x+5.24	0.97
True	y=-1.48x+3.75	0.94	y=-2.49x+5.67	0.95

TABLE.VI.
Evaluate criteria using *D*-statistic

	Dolphin2				iPlane			
m/k	dd	cd	spld	AVG	dd	cd	spld	AVG
2	**0.331**	5.1E-3	0.258	0.198	**0.131**	2.9E-4	0.249	0.128
8	0.359	1.3E-3	0.182	**0.181**	0.164	6.0E-5	**0.145**	**0.103**
26	0.385	6.2E-3	**0.144**	0.178	0.180	9.0E-5	0.180	0.120
106	0.461	**1.0E-3**	0.198	0.220	0.234	2.0E-5	0.286	0.173
425	0.453	2.0E-3	0.176	0.210	0.278	**1.0E-5**	0.270	0.183

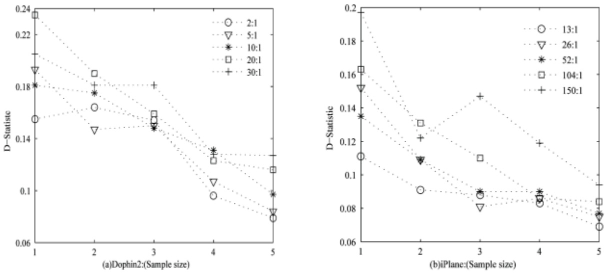


Fig.3. The average *D*-statistic of the *D*-statistic of *dd*, *cd*, *spld* as the sample size varies.

Using the two snapshots of the router-level topology, we make the same measurements with the ones on the simulated graphs. In matching the graph pattern, we set 5 values of *m/k* ranging from 2 to 425 and take 1 percent of the routes for each *m/k* 10 times. Table V shows the trend lines of one representative result. In comparing graph patterns using the *D*-statistic, we set *m/k* the same as above and each entry in table VI is obtained by averaging 10 different sample sizes ranging from 1 to 5 percent. In evaluating the scalability of *m/k*, we set another two groups of *m/k* because of the constriction of the number of sources and destinations of the data sets, and five different sample sizes ranging from 1 to 16 percent of the routes. Figure 3 shows how the quality of the sample graph changes with the sample size for each *m/k*.

From these results, we can derive the qualitatively observations: Both on the two snapshots of the router-level topology, the smallest *m/k* performs best on estimating the exponent and *dd*, and the accuracy of the two properties will decreases as *m/k* increases (see table V and table VI). For the overall of the sample graph, the small *m/k* often performs better than the large one even when the sample size increases to 16 percent of the routes (see figure 3).

Another conclusion is that there is no single *m/k* answer to all the properties we test. In summary, results on the two snapshots are consistent with the simulated graphs, especially the results we obtained on the EBA graph. Thus we can reasonably argue the general method applied in many famous topology measurement systems that taking plentiful destinations while a small number of sources may cause the observation of properties of graph differ sharply from the underlying graph.

Note that though the two data sets we choose are currently the most plentiful results of the IPv4 and IPv6 networks, they are also obtained by traceroute which have been proved biased and incomplete. Another potential factor influencing our result is that we take the same sources and destinations of the real exploration in order to simulate the scenario of *(k,m)-traceroute* exploration on the real-world network which restricts the randomness of the sample result. This can also cause the evaluation of *m/k* biased. But because the sample we take is small (all sizes of the samples are no more than 16 percent of the raw traces) compared to the original map which promises the randomness of the sample, we believe our result is credible.

V. Conclusion and Futrue Work

In this study, we conducted an extensive set of simulations to study the impact of a variety of proportions of destinations and sources on estimating the Internet topology in perspective of several properties. To achieve this, we took the most commonly used models (they are the ER, the BA and EBA models). Our results show that different proportions of destinations and sources have a significant impact on the accuracy of the sample graphs even if the numbers of total probes are the same. In particular, these results imply that the general method of taking a small set of sources to perform traceroute-like probes to a large set of destinations attempting to obtain a more accurate topology of the Internet is very unreasonable which could make the properties of graph differ sharply from the underlying graph. It also suggests that there seems to be no single perfect *m/k* answer to the graph sampling and the small *m/k* often perform better than the large overall. We also validate this conclusion on the real-world networks.

Currently, we are doing more investigations. First, we are considering more models, like DM[26] and GL[27], that with high clustering, scaling the simulated graphs up to a large size and checking more properties. From figure 2 and figure 3, we see that the sample size is a more important factor than the *m/k* to the impact on the graph properties and we are also studying the reason. Another related field is the correlation between the network scale and the *m/k* which will become interesting as the IPv6 network grows. Note that what we concern in this paper is to present the correlation between different proportions of destinations and sources and the accuracy of graph properties and prove *m/k*'s significant impact on the accuracy of the graph properties, but not to find the nature hidden in the phenomena, e.g., how different *m/k* influence the properties of graph.

Acknowledgements

The research was supported by National 973 Program of China (Grant No.2005CB321901) and the fund of the State Key Laboratory of Software Development Environment (SKLSDE-2008ZX-03). We thank *CAIDA* and *iPlane* for providing us with the data sets.

References

[1] CAIDA Skitter tool, http://www.caida.org/tools/measurement/skitter/.

[2] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling Biases in IP Topology Measurements. In IEEE INFOCOM, 2003.

[3] J.-L. Guillaume, M. Latapy. Relevance of Massively Distributed Explorations of the Internet Topology: Simulation Results. In IEEE INFOCOM, 2005.

[4] R. Govindan and H. Tangmunarunkit. Heuristics for Internet Map Discovery. In IEEE INFOCOM, Mar. 2000.

[5] D. G. Waddington, F.-Z. Chang, R. Viswanathan, B. Yao. Topology Discovery for Public IPv6 Networks. In ACM SIGCOMM Computer Communication Review, vol.33, no. 3, July 2003, pp. 59–68.

[6] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP Topologies with Rocketfuel. In ACM SIGCOMM, Aug 2002.

[7] R. Albert and A.-L. Barabási. Statistical Mechanics of Complex Network. Reviews of Modern Physics 74,2002,47~97.

[8] R. Cohen, K. Erez, D. ben-Avraham, S. Havlin. Resilience of the Internet to Random Breakdown. Phys. Rev. Lett., 86:4626-4628,2000.

[9] D. Magoni and J.-J. Pansiot. Influence of Network Topology on Protocol Simulation. In ICN'01-1st IEEE International Conference on Networking, volume Lecture Notes in Computer Science, pages 762.770, July 9-13, 2001.

[10] A. Clauset and C. Moore. Traceroute Sampling Makes Random Graphs Appear to Have Power Law Degree Distributions. cond-mat/0312674.

[11] P. De Los Rios. Exploration Bias of Complex Networks. The 7th Conference on Statistical and Computational Physics Granada, 2002.

[12] Cooperative Association for Internet Data Analysis. http://www.caida.org/.

[13] DIMES@home Project. http://www.netdimes.org/new/.

[14] Traceroute@Home project. University of Paris 6, coordinator: Timurfriedman.

[15] B. Donnet, P. Raoult, T. Friedman. Deployment of an Algorithm for Large-Scale Topology Discovery. In Proceedings of ACM SIGMETRICS, 2005.

[16] B. Bollobás. Random Graphs. Academic Press, 1985.

[17] R. Albert and A.-L. Barabási. Emergence of Scaling in Random Networks. Science, 286:509-512,1999.

[18] R. Albert and A.-L. Barabási. Topology of Evolving Networks: Local Events and Universality. Physical Review Letters, 2000.

[19] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. Exploring Networks with Traceroute-Like Probes: Theory and simulations. Theoretical Computer Science, Special Issue on Complex Networks, 2005.

[20] H. Haddadi, M. Rio, G. Iannaccone, A. Moore, R. Mortier. Network Topologies: Inference, Modeling, and Generation. IEEE Communications Surveys & Tutorials 2nd Quarter 2008.

[21] http://www.physics.csbsju.edu/stats/KS-test.html

[22] http://www.labri.fr/perso/magoni/nem/

[23] X. Zou, Z.-L. Qiao, G. Zhou, K. Xu. A Logic Distance-Based Method for Deploying Probing Sources in the Topology Discovery. IEEE GLOBECOM 2009.

[24] http://ipv6.nlsde.buaa.edu.cn/

[25] http://iplane.cs.washington.edu/

[26] S.N. Dorogovtsev and J.F.F.Mendes. Evolution of networks. Adv.Phys.51,1079-1187,2002.

[27] J.-L.Guillaume and M.Latapy. Bipartite graphs as models of complex networks. Physica A: Statistical and Theoretical Physics Volume 371, Issue 2, 15 November 2006, Pages 795-813.