How accurate are `traceroute`*-like Internet mappings ?*

Luca Dall'Asta[1], Ignacio Alvarez-Hamelin[1,4], Alain Barrat[1], Alexei Vázquez[2] and Alessandro Vespignani[3]

[1]*Laboratoire de Physique Théorique, Bâtiment 210, Université de Paris-Sud, 91405 ORSAY Cedex France*
[2]*Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN 46556, USA.*
[3]*School of Informatics and Department of Physics, University of Indiana, Bloomington, IN 47408, USA*
[4] *Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, C 1063 ACV Buenos Aires, Argentina*

Mapping the Internet generally consists in sampling the network from a limited set of sources by using `traceroute`-like probes. This methodology, akin to the merging of different spanning trees to a set of destinations, has been argued to introduce uncontrolled sampling biases that might produce statistical properties of the sampled graph which sharply differ from the original ones[1, 2, 3]. In this paper we study numerically how the fraction of vertices and edges discovered in the sampled graph depends on the particular deployments of probing sources. The results might hint the steps toward more efficient mapping strategies.

Keywords: Traceroute, Internet exploration, Topology inference

1 Introduction

In the absence of accurate Internet maps, researchers rely on a general strategy that consists in acquiring local views of the network from several vantage points and merging these views in order to get a presumably accurate global map. By using this strategy, a number of research groups have generated maps of the Internet [4, 5, 6, 7, 8], that have been used for the statistical characterization of the network properties. Defining $\mathcal{G} = (V,E)$ as the sampled graph of the Internet with $N = |V|$ vertices and $|E|$ edges, it is quite intuitive that the Internet is a *sparse* graph in which the number of edges is much lower than in a complete graph; i.e. $|E| \ll N(N-1)/2$. Equally important is the fact that the average distance, measured as the shortest path, between vertices is very small. This is the so called *small-world* property, that is essential for the efficient functioning of the network. Most surprising is the evidence of a skewed and heavy-tailed behavior for the probability that any vertex in the graph has degree k defined as the number of edges linking each vertex to its neighbors. In particular, in several instances, the degree distribution appears to be approximated by $P(k) \sim k^{-\gamma}$ with $2 \leq \gamma \leq 2.5$ [9]. Evidence for the heavy-tailed behavior of the degree distribution has been collected in several other studies at the router and AS level [10, 11, 12, 13, 14] and have generated a large activity in the field of network modeling and characterization [15, 16, 17, 18, 19].

While `traceroute`-driven strategies are very flexible and can be feasible for extensive use, the obtained maps are undoubtedly incomplete. Along with technical problems such as the instability of paths between routers and interface resolutions [20], typical mapping projects are run from relatively small sets of sources whose combined views are missing a considerable number of edges and vertices [14, 21]. In particular, the various spanning trees are specially missing the lateral connectivity of targets and sample more frequently vertices and links which are closer to each source, introducing spurious effects that might seriously compromise the statistical accuracy of the sampled graph. These *sampling biases* have been explored in numerical experiments of synthetic graphs generated by different algorithms[1, 2, 3, 24].

It was shown in [22] that the map accuracy depends on the underlying network *betweenness centrality* [†] distribution. We substantiate the analytical finding of [22] with a throughout exploration of maps obtained varying the number of source-target pairs on networks models with different topological properties.

2 Optimization of mapping strategies

Let us consider sparse undirected graphs denoted by $G = (V, E)$. In particular, we will consider two main classes of graphs: *i) Homogeneous graphs* in which the degree distribution $P(k)$ has small fluctuations and a well defined average degree; *ii) Heterogeneous graphs* for which $P(k)$ is a broad distribution with heavy-tail and large fluctuations.

The most widely known model for homogeneous graphs is given by the classical Erdös-Rényi (ER) model [23]: in such random graphs $G_{N,p}$ of N vertices, each edge is present in E independently with probability p. We generated ER graphs with $p = 1/N$, where $N = 10^4$.

In opposition to the previous case, heterogeneous graphs are characterized by connectivity distributions spanning various orders of magnitude, with a heavy-tail at large k. While we do not want to enter the detailed definition of heavy-tailed distribution we have considered two classes of such distributions: (i) *scale-free* or Pareto distributions of the form $P(k) \sim k^{-\gamma}$ (RSF), and (ii) Weibull distributions (WEI) $P(k) = (a/c)(k/c)^{a-1}\exp(-(k/c)^a)$. In both cases, we have generated the corresponding random graphs by using the algorithm proposed by Molloy and Reed [25]. The parameters used are $a = 0.25$ and $c = 0.6$ for the Weibull distribution, and $\gamma = 2.3$ for the RSF case, and all graphs have $N = 10^4$ nodes.

It was shown in [22] that it is possible to have a general qualitative understanding of the efficiency of network exploration and the induced biases on the statistical properties. The quantitative analysis of the sampling strategies, however, is a much harder task that calls for a detailed study of the discovered proportion of the underlying graph and the precise deployment of sources and targets. In this perspective, very important quantities are the fraction N^*/N and E^*/E of vertices [‡] and edges discovered in the sampled graph, respectively. In our study the parameters of interest are the density $\rho_T = N_T/N$ and $\rho_S = N_S/N$ of targets and sources. An appropriate quantity representing the level of sampling of the networks is $\varepsilon = \frac{N_S N_T}{N}$, that measures the density of probes imposed to the system.

This finding hints toward a behavior that is determined by the number of sources and targets, N_S and N_T. Any quantity is thus a function of N_S and N_T, or equivalently of N_S and ρ_T. This point is clearly illustrated in Fig. 1, where we report the behavior of E^*/E and N^*/N at fixed ε and varying N_S and ρ_T. The curves exhibit a non-trivial behavior and since we will work at fixed $\varepsilon = \rho_T N_S$, any measured quantity can then be written as $f(\rho_T, \varepsilon/\rho_T) = g_\varepsilon(\rho_T)$. Very interestingly, the curves show a structure allowing for local minima and maxima in the discovered portion of the underlying graph.

This feature can be explained by a simple symmetry argument. The model for `traceroute` is symmetric by the exchange of sources and targets, which are the endpoints of shortest paths: an exploration with $(N_T, N_S) = (N_1, N_2)$ is equivalent to one with $(N_T, N_S) = (N_2, N_1)$. In other words, at fixed $\varepsilon = N_1 N_2/N$, a density of targets $\rho_T = N_1/N$ is equivalent to a density $\rho'_T = N_2/N$. Since $N_2 = \varepsilon/\rho_T$ we obtain that at constant ε, experiments with ρ_T and $\rho'_T = \varepsilon/(N\rho_T)$ are equivalent obtaining by symmetry that any measured quantity obeys the equality $g_\varepsilon(\rho_T) = g_\varepsilon\left(\frac{\varepsilon}{N\rho_T}\right)$. This relation implies a symmetry point signaling the presence of a maximum or a minimum at $\rho_T = \varepsilon/(N\rho_T)$. We therefore expect the occurrence of a symmetry in the graphs of Fig.1 at $\rho_T \simeq \sqrt{\varepsilon/N}$. Indeed, the symmetry point is clearly visible and in quantitative good agreement with the previous estimate in the case of heterogeneous graphs. On the contrary, homogeneous underlying topology have a smooth behavior that makes difficult the clear identification of the symmetry point. Moreover, unique shortest path probes create a certain level of correlations in the exploration that tends to hide the complete symmetry of the curves.

The previous results imply that at fixed levels of probing ε different proportions of sources and targets may achieve different levels of sampling. This hints to the search for optimal strategies in the relative

[†] The betweenness represents the all-to-all traffic situation.

[‡] The measured quantities have the symbol $*$, to distinguish from the original ones.

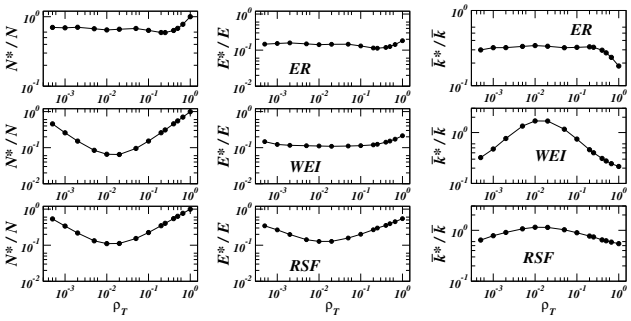

Fig. 1: Behavior as a function of ρ_T of the fraction of discovered edges and vertices in explorations with fixed ε (here $\varepsilon = 2$). Since $\varepsilon = \rho_T N_S$, the increase of ρ_T corresponds to a lowering of the number of sources N_S. The plots on the right show the fraction of the normalized average degree $\overline{k}^*/\overline{k}$.

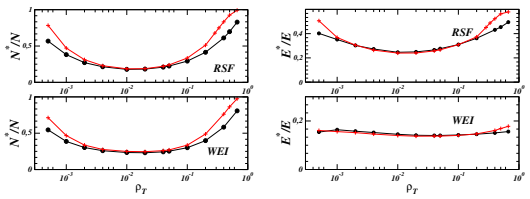

Fig. 2: Behavior as a function of ρ_T of the fraction of discovered edges and vertices in explorations with fixed ε (here $\varepsilon = 2$). The circles correspond to a random deployment of sources and targets while the crosses are obtained when sources and targets are vertices with lowest betweenness vertices.

deployment of sources and targets. The picture, however, is more complicate if we look at other quantities in the sampled graph. In Fig.1 we show the behavior at fixed ε of the average degree $\overline{k}^*$ measured in sampled graphs normalized by the actual average degree $\overline{k}$ of the underlying graph as a function of ρ_T. The plot shows also in this case a symmetric structure. By comparing the data of Fig.1 we notice that the symmetry point is of a different nature for different quantities: the minimum in the fraction of discovered edges corresponds to the best estimate of the average degree. In other words, the best level of sampling is achieved at particular values of ε and N_S that are conflicting with the best sampling of other quantities.

The evidence purported in this section hints to a possible optimization of the sampling strategy. The optimal solution, however, appears as a trade-off strategy between the different level of efficiency achieved in competing ranges of the experimental setup. In this respect, a detailed and quantitative investigation of the various quantities of interest in different experimental setups is needed in order to pinpoint the most efficient deployment of source-target pairs depending on the underlying graph topology. While such a detailed analysis lies beyond the scope of the present study, an interesting hint comes from the analytical results of [22]: since vertices with large betweenness have typically a very large probability of being discovered, placing the sources and targets preferentially on low-betweenness vertices (the most difficult to discover) may have an impact on the whole process. This is what we investigate in Fig. 2 in which we report the fraction of vertices and edges discovered by either a random deployment of sources and targets or a deployment on the lowest-betweenness vertices. It is apparent that such a deployment allows to discover larger parts of the network. Of course the procedure used is unrealistic since identifying low-betweenness vertices is not an easy task. The usual correlation between connectivity and betweenness however indicates that the exploration of a real network could be improved by a massive deployment of sources using low-connectivity vertices.

3 Conclusions and outlook

The rationalization of the exploration biases at the statistical level provides a general interpretative framework for the results obtained from the numerical experiments on graph models. In general, exploration strategies provide sampled distributions with enough signatures to distinguish at the statistical level between graphs with different topologies. It is of major importance to define strategies that optimize the estimate of the various parameters and quantities of the underlying graph. In this paper we have shown that the proportion of sources and targets may have an impact on the accuracy of the measurements even if the number of total probes imposed to the system is the same. For instance, the deployment of a highly distributed infrastructure of sources probing a limited number of targets may result as efficient as few very powerful sources probing a large fraction of the addressable space [26]. The optimization of large network sampling is therefore an open problem that calls for further work aimed at a more quantitative assessment of the mapping strategies both on the analytic and numerical side.

References

[1] A. Lakhina, J. W. Byers, M. Crovella and P. Xie, "Sampling Biases in IP Topology Measurements," Technical Report BUCS-TR-2002-021, Department of Computer Sciences, Boston University (2002).

[2] A. Clauset and C. Moore, "Accuracy and Scaling Phenomena in Internet Mapping," Phys. Rev. Lett. **94**, 018701 (2005).

[3] T. Petermann and P. De Los Rios, "Exploration of Scale-Free Networks - Do we measure the real exponents?," Eur. Phys. J. B **38** 201-204 (2004).

[4] The National Laboratory for Applied Network Research (NLANR), sponsored by the National Science Foundation. (see http://moat.nlanr.net/).

[5] The Cooperative Association for Internet Data Analysis (CAIDA), located at the San Diego Supercomputer Center. (see http://www.caida.org/home/).

[6] Topology project, Electric Engineering and Computer Science Department, University of Michigan (http://topology.eecs.umich.edu/).

[7] SCAN project at the Information Sciences Institute (http://www.isi.edu/div7/scan/).

[8] Internet mapping project at Lucent Bell Labs (http://www.cs.bell-labs.com/who/ches/map/).

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On Power-law Relationships of the Internet Topology," ACM SIGCOMM '99, Comput. Commun. Rev. **29**, 251–262 (1999).

[10] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet Map Discovery," Proc. of IEEE Infocom 2000, Volume 3, IEEE Computer Society Press, 1371–1380, (2000).

[11] A. Broido and K. C. Claffy, "Internet topology: connectivity of IP graphs," San Diego Proceedings of SPIE International symposium on Convergence of IT and Communication. Denver, CO. 2001

[12] G. Caldarelli, R. Marchetti, and L. Pietronero, "The Fractal Properties of Internet," Europhys. Lett. **52**, 386 (2000).

[13] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, "Dynamical and Correlation Properties of the Internet," Phys. Rev. Lett. **87**, 258701 (2001); A. Vázquez, R. Pastor-Satorras, and A. Vespignani, "Large-scale topological and dynamical properties of the Internet," Phys. Rev. E .**65**, 066130 (2002).

[14] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, "The Origin of Power Laws in Internet Topologies Revisited," Proceedings of IEEE Infocom 2002, New York, USA.

[15] A. Medina and I. Matta, "BRITE: a flexible generator of Internet topologies," Tech. Rep. BU-CS-TR-2000-005, Boston University, 2000.

[16] C. Jin, Q. Chen, and S. Jamin, "INET: Internet topology generators," Tech. Rep. CSE-TR-433-00, EECS Dept., University of Michigan, 2000.

[17] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[18] P.Baldi, P.Frasconi and P.Smyth, *Modeling the Internet and the Web: Probabilistic methods and algorithms*(Wiley, Chichester, 2003).

[19] R. Pastor-Satorras and A. Vespignani, *Evolution and structure of the Internet: A statistical physics approach* (Cambridge University Press, Cambridge, 2004).

[20] H. Burch and B. Cheswick, "Mapping the internet," IEEE computer, **32**(**4**), 97–98 (1999).

[21] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker, "Scaling phenomena in the Internet: Critically examining criticality," Proc. Natl. Acad. Sci USA **99** 2573–2580, (2002).

[22] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vzquez, A. Vespignani "Traceroute-like exploration of unknown networks: a statistical analysis" in Proc of Combinatorial and Algorithmic Aspects of Networking and the Internet August 5 - 7, 2004, Banff, Canada, to apear in LCNS.

[23] P. Erdös and P. Rényi, "On random graphs I," Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[24] J.-L. Guillaume and M. Latapy, "Relevance of Massively Distributed Explorations of the Internet Topology: Simulation Results," Proc. Infocom 2005 (to appear).

[25] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," Random Struct. Algorithms **6**, 161 (1995). M. Molloy and B. Reed, "The size of the giant component of a random graph with a given degree distribution," Combinatorics, Probab. Comput. **7**, 295 (1998).

[26] http://www.tracerouteathome.net/