Network sampling: effect of realistic modeling of the traceroute tool

A. Barrat and A. Vespignani

2007
Network sampling: effect of realistic modeling of the traceroute tool

Alain Barrat,∗ Alessandro Vespignani†

1 Introduction

A significant research and technical challenge in the study of large information networks is related to the lack of highly accurate maps providing information on their basic topology. This is mainly due to the dynamical nature of their structure and to the lack of any centralized control resulting in a self-organized growth and evolution of these systems. A prototypical example of this situation is faced in the case of the physical Internet. The topology of the Internet can be investigated at different granularity levels such as the router and Autonomous System (AS) level, with the final aim of obtaining an abstract representation where the set of routers (ASs) and their physical connections (peering relations) are the vertices and edges of a graph, respectively. In the absence of accurate maps, local views are obtained by evaluating a certain number of paths to different destinations by using specific tools such as traceroute or by the analysis of BGP tables. At first approximation these processes amount to the collection of paths from a source vertex to a set of target vertices, obtaining a partial spanning tree of the network. The merging of several of these views provides the map of the Internet from which the statistical properties of the network are evaluated.

This strategy has led to the obtention of various maps of the Internet [1, 2, 3, 4, 5, 6] which have been used for the statistical characterization of the network. Defining \( G = (V, E) \) 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 with a much lower number of edges than in a complete graph: \( |E| \ll N(N - 1)/2 \). Moreover, 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, the degree distribution appears to be approximated by \( P(k) \sim k^{-\gamma} \) with \( 2 \leq \gamma \leq 2.5 \) [7]. Evidence for the heavy-tailed behavior of the degree distribution has been collected in several other studies at the router and AS level [14, 15, 16, 17, 18] and have generated a large activity in the field of network modeling and characterization [19, 8, 9, 10, 11].

The obtained maps are however undoubtedly incomplete. Along with technical problems such as the instability of paths between routers and interface resolutions [12], typical mapping projects are run from relatively small sets of sources whose combined views are missing a considerable number of edges and vertices [18, 13]. 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 [20, 21, 26, 22, 29, 30]. In particular, it has been shown (numerically and analytically) that apparent degree distributions with heavy-tails may be observed even from homogeneous topologies such as in the classic Erdős-Rényi

∗Université de Paris-Sud, France, alain.barrat@th.u-psud.fr
†Indiana University, USA, and Université de Paris-Sud, France
graph model [20, 21, 23]. These initial studies thus pointed out that the evidence obtained from the analysis of the Internet sampled graphs might be insufficient to draw conclusions on the topology of the actual Internet network.

In order to shed more light on this delicate issue, sampling biases need to be explored systematically. The method consists in mimicking the sampling by traceroute on known graphs with various topological characteristics, and to compare the properties of the sampled graph with the original one [22, 29, 30]. The crucial step lies in the modeling of the sampling procedure, i.e. of the traceroute tool. Up to now, studies have relied to the following simple scheme: a certain number of sources and targets are selected randomly ¹, and the sampling amounts to the discovery of the shortest paths between each source-target pair. This simple assumption considers therefore that traceroute probes follow the topological shortest path between source and target, and neglects the many factors, including commercial agreement, traffic congestion and administrative routing policies, that contribute to determine the actual path. Despite these local, often unpredictable path distortions, the shortest path seems indeed to yield a reasonable first approximation of the route traversed by traceroute-like probes. Moreover, such an approach lends itself to an analytical approximation which allows to grasp the origin of the sampling biases and to connect them to the topological properties of the network [29, 30] ². These analysis show that, at a qualitative level, the main conclusions drawn from traceroute-like samplings are reliable. In particular, such samplings allow for accurate discrimination between topologies with degree distributions that are heavy-tailed from those that are homogeneous [29, 30], as soon as the sampling effort is not reduced to a single source (as in [23, 21]). On the other hand, at a quantitative level even considerable deviations between numerical summaries of characteristics of the sampled networks and those of the actual Internet are possible.

In order to go further in assessing reliability and efficiency of traceroute sampling, and in understanding the sampling biases, the use of a more realistic model for the paths followed by traceroute probes is necessary. Leguay et al. [24] have used data from real Internet sampling experiments to compare in details the real paths followed by the probes with the shortest paths, and found for example that the former are systematically longer than the latter (at least 80% of the paths are not shortest paths). It is computationally very difficult to sample randomly paths between a source $s$ and a target $t$ of length $\ell + \delta$ where $\ell$ is the shortest path length and $\delta$ an additional length which almost does not depend on $\ell$. A simple “Random Deviation” model was therefore put forward, in order to reproduce most paths characteristics [24]. In this report, we use the Random Deviation Model of [24] to perform numerical experiments of traceroute like sampling, in order to reconsider the results of [22, 29, 30] and to understand how they are modified by the use of a more realistic sampling model.

The report is organized as follows: we first present the detailed model and the networks investigated in section 2; we show in section 3 how the ability to separate homogeneous from heterogeneous networks topologies is preserved, while quantitative biases are investigated in section 4.

2 The model

In a typical traceroute study, a set of active traceroute sources deployed in the network run traceroute probes to a set of destination nodes. Each probe collects information on all the nodes and edges traversed along the path connecting the source to the destination, allowing the discovery of the network. By merging the information collected on each path it is then possible to reconstruct a partial map of the network. We therefore model the sampling process by the random placement of $N_S$ sources and $N_T$ destination targets. The parameters of interest are the density $\rho_T = N_T/N$ and $\rho_S = N_S/N$ of targets

¹Non-random choices have however also been briefly investigated [22, 29, 30]

²In this framework, extensive numerical experiments are also easily performed for a large variety of synthetic networks [22, 29, 30]
and sources. In general, traceroute-driven studies run from a relatively small number of sources to a much larger set of destinations. For this reason, in many cases it is appropriate to work with the density of targets \( \rho_T \) while still considering \( N_S \) instead of the corresponding density. Indeed, it is clear that while 100 targets may represent a fair probing of a network composed by 500 vertices, this number would be clearly inadequate in a network of \( 10^6 \) vertices. On the contrary, the density of targets \( \rho_T \) allows us to compare mapping processes on networks with different sizes by defining an intrinsic percentage of targeted vertices. In many cases, an appropriate quantity representing the level of sampling of the networks is \( \epsilon = N_S N_T / N \), that measures the density of probes imposed to the system. In real situations it represents the density of traceroute probes in the network and therefore a measure of the load provided to the network by the measuring infrastructure.

The next modeling step is to define, for each source-target pair, and therefore for each traceroute sent from a source towards a target, which intermediate nodes are discovered. As mentioned in the introduction, previous studies have considered that probes travel along the shortest paths (SP) connecting each source-target pair [23, 21, 22, 29, 30].

In the real Internet, many factors, including commercial agreement, traffic congestion and administrative routing policies, contribute to determine the actual path, causing it to differ from the shortest path. It seems very difficult to precisely model these local, often unpredictable path distortions or inflations, but, in order to go beyond the first approximation of shortest paths, Ref [24] proposed an effective way to take into account the traceroute paths distortions. The following more realistic path selection criterion was put forward for a data packet sent from a source \( S \) to a target \( T \): at each step, the packet moves with probability \( 1 - p \) towards \( T \) along the shortest path from its current position to \( T \), and it performs instead a step in a random direction with probability \( p \). Clearly, this Random Deviation Model (RDM) lengthen the paths between source and destination, and the authors of [24] show that a probability \( p \sim 0.2 \) allows to accurately reproduce most properties of the true traceroute-paths followed, and in particular the distribution length. Of course, real paths are not random, but the large number of uncontrolled factors shaping them can be considered as statistically equivalent to such random perturbations.

The traceroute sampling is thus modeled by computing, for each ensemble of source-target pairs \( \Omega = \{ S, T \} \), paths connecting each source-target pair along these lines. The schematic procedure is illustrated in figure 1. All vertices and links belonging to the chosen paths are considered as discovered by the sampling process, and the sampled graph \( G = (V^*, E^*) \) is defined as the set of vertices \( V^* \) (with \( N^* = |V^*| \)) and edges \( E^* \) induced by considering the union of all the paths connecting the source-target pairs. Clearly, for a given number a sources and targets, and therefore of data packets sent, a non-zero amount of randomness in the routing of packets will lead to the discovery of longer paths and therefore of larger portions of the network. We will investigate this issue in more details in the next sections, by comparing the sampling biases introduced in the characteristics of networks with different topologies by traceroute sampling using either shortest paths (SP) or shortest paths with random deviations (RDM) between sources and targets. As in previous studies, we will focus on underlying graphs with different topologies:

A) Homogeneous graphs in which the degree distribution \( P(k) \) has small fluctuations and a well defined average degree. In this context, the homogeneity refers to the existence of a meaningful characteristic average degree that represents the typical value in the graph. We will consider the most widely known model for homogeneous graphs, namely the classical Erdős-Rényi (ER) model [25].

B) Heterogeneous graphs for which \( P(k) \) is a broad distribution with heavy-tail and large fluctuations, spanning various orders of magnitude. We will consider the well-known Barabási-Albert model [27], as well as the scale-free graph with large clustering introduced by Dorogovtsev, Mendes and Samukhin [28], and intermediate growing networks with tunable clustering coefficient. We will use a network with clustering coefficient \( C \sim 0.2 \), similar results being obtained with other values of \( C \).
Figure 1: Schematic illustration of the models for traceroute. The discovered nodes and links are shown by filled circles and continuous lines, respectively. Undiscovered elements are shown by empty circles and dashed lines. Top: the packets sent by the source to the target follow the shortest path (SP). Bottom: Random Deviation Model in which, at each step, the packet undergoes a random deviation with a certain probability (here in A and B), and continues afterwards on the new shortest path.

3 Global picture, degree distributions

A large part of the debate on sampling biases introduced through sampling of networks was centered about the heterogeneous character of the degree distribution. It has indeed been shown (numerically and analytically) that apparent degree distributions with heavy-tails may be observed even from homogeneous topologies such as in the classic Erdős-Rényi graph model [20, 21, 23]. In a reassuring way however, subsequent studies [29, 30] pointed out that this strong bias occurs only when the sampling effort is reduced to only one source, which is not the case in the real Internet sampling experiments. Moreover, as soon as the sampling effort is increased, the sampled graph clearly distinguishes the two situations defined by homogeneous and heavy-tailed topologies, respectively. This is due to the exploration process that statistically focuses on high betweenness nodes, thus providing a very accurate sampling of the distribution tail. In graphs with heavy-tails, such as scale-free networks, the main topological features are therefore easily discriminated since the relevant statistical information is encapsulated in the degree distribution tail which is fairly well captured.

While the cited studies were done using the SP model for traceroute, the use of more sophisticated models for traceroute could a priori alter at least slightly these qualitative findings. Taking into account random deviations in the paths connecting sources and targets lead indeed intuitively to the discovery of more nodes and links, but one could imagine that this process introduces new uncontrolled biases.

A first insight is given by the fact that the paths followed by the sampling process are a mixture of shortest paths and random walks. If only shortest paths are discovered, the betweenness centrality of nodes and links directly determine their probability to be discovered [29, 30]. The mean-field analysis of [29, 30] can in fact be performed for any path selection criterion: the probability of discovery is then linked
to a modified betweenness centrality which measures how many paths go through each node or link. If the path were totally random, the adequate betweenness measure would be the random walk betweenness centrality [31]. In most networks, both shortest path betweenness and random walk betweenness are strongly correlated with the degree. In both cases therefore, and also in the case of RDM paths which combine shortest paths and randomness, the path-based sampling will yield large discovery probabilities for the large degree nodes, and the qualitative picture obtained in [29, 30] should remain valid.

Figure 2: Degree distribution of sampled networks, using either shortest-paths (SP) or the random deviation model (RDM) to model the sampling by traceroute. The underlying networks are A) an ER graph with $\langle k \rangle = 60$ and B) a BA graph with degree distribution $P(k) \sim k^{-3}$.

The numerical experiments indeed confirm these ideas. Figure 2 display the sampled degree distributions obtained for traceroute-like sampling, with SP and RDM, of homogeneous (ER) and heterogeneous (BA) networks. As already noted in [29, 30], the sampling of homogeneous networks is somewhat cumbersome: the original distribution is peaked around its average $\langle k \rangle$, but the sampled distributions span a wider range of degrees, from 1 to a cutoff close to $\langle k \rangle$ (nodes cannot be discovered with a larger degree than their real one). No power-law shape is however obtained. When the underlying network is heterogeneous, with a broad degree distribution, the sampled distribution is broad as well with both sampling models. Moreover, for a given sampling set-up, i.e. for given numbers of sources and targets $N_S$ and $N_T$, RDM sampling leads to larger cutoffs than SP sampling. This observation can be understood in the following way: as previously mentioned, the hubs have large betweenness centrality and many sampling paths go through them. The various paths could however arrive to and depart from the hubs through a restricted number of links (those with largest betweenness) and ignore many connections, thus reducing strongly the discovered degree of the nodes. The possibility introduced in the RDM, to perform a step at random instead of following the shortest path, clearly increases strongly the probability to discover more links when passing through a node, as shown schematically in Figure 3: let us consider two sources $s_1$ and $s_2$ and a target $t$, such that both SP from sources to $t$ go through a hub $i$. If the data packet sent from $s_1$ to $t$ follows the shortest path through $i$, while the packet sent from $s_2$ undergoes a random deviation while passing through $i$, an additional link going out of $i$ will be discovered with respect to the situation in which all packets follow shortest paths. This effect is present for all nodes, but hubs through which many paths go will be particularly sensitive to it. The discovered degree of hubs will thus be enhanced by the more realistic sampling, leading to an even better sampling of the degree distribution tail.

4 Quantitative effects

Let us now investigate how the fraction of network discovered depends on the sampling effort. In particular, we measure the fraction of nodes $N^*/N$ and links $E^*/E$ discovered, as well as the ratio between the clus-
Figure 3: Illustration of the enhancement by RDM of the degree of discovered nodes. The discovered nodes and links are shown by filled circles and continuous lines, respectively. Undiscovered elements are shown by empty circles and dashed lines. For SP sampling, the two paths going through \( i \) discover the same links and nodes between \( i \) and the target. For RDM, in the case of a random deviation in \( i \), a new link connected to \( i \) is discovered.

Figure 4: Comparison of the discovery ratios for SP and RDM, as a function of the target density \( \rho_T \), for various networks, namely the ER homogeneous network, the BA scale-free network with small clustering, and the BA-DM scale-free network with stronger clustering properties. Data are for sampling efforts \( \epsilon = 2 \) and \( \epsilon = 5 \), and averaged (for each network type) over 20 sampling realizations on each of 10 different networks of \( N = 10000 \) nodes.

A very interesting point concerns the clustering coefficient. As illustrated in figure 5, the RDM allows to discover triangles more easily than the use of SP. In fact, figure 4 shows that the clustering coefficient of
the sampled graph is even more enhanced by RDM with respect to SP than the other characteristics ($N^*$ or $E^*$). This increase in $C^*$ is therefore not only due to the fact that more nodes and links are discovered but rather to the structural difference between the paths followed by the probes. Interestingly, the ratio $C^*/C$ can even exceed 1 for networks with small clustering (ER and BA), leading to a possible overestimation of the clustering; the order of magnitude remains however the same: the sampling of a network with very small clustering cannot yield a sampled network with a clustering coefficient of order 1 as observed for the Internet. For clustered networks, $C^*/C$ remains remains smaller than 1, and values quite close to 1 are observed for RDM, in contrast with SP which leads to strong underestimates of the clustering. This result allows to rationalize the data of real Internet sampling experiments which indeed lead to large clustering coefficients of order 0.5, while sampling through SP of even strongly clustered networks (e.g. $C \sim 0.8$) yield typically sampled values of 0.1.

![Diagram](image-url)

Figure 5: Illustration of the enhancement by RDM of the degree of discovered nodes. The discovered nodes and links are shown by filled circles and continuous lines, respectively. Undiscovered elements are shown by empty circles and dashed lines. For SP sampling, the paths from the three sources to the target would allow to discover only the links $ik$ and $jk$, and no shortest path to the target goes through $ij$, while with RDM, the third link $ij$ has some probability to be discovered, as well as the triangle $ijk$.

A more detailed insight into this issue is obtained by the investigation of the clustering spectra of the sampled and original networks, denoted respectively by $C^*(k)$ and $C(k)$. Figure 6 shows that RDM yields a spectrum closer to the original $C(k)$ than SP for the clustered network. For both clustered and non-clustered networks, RDM leads to a strong increase of $C^*(k)$ with respect to SP especially at small $k$. This can be understood thanks to the analysis of the sampling processes already developed: the triangles around the hubs are more easily discovered (since the hubs themselves are well discovered [30]) even with SP sampling. On the other hand, triangles around nodes of small degree typically contain nodes and links of small betweenness, and some links may be discovered only if some random deviations send the probes through them. From the clustering point of view, the effect of random deviations is thus stronger for small degree nodes.

Let us finally investigate the effect of the various traceroute-like samplings on the $k$-core decomposition of the networks. The $k$-core decomposition [32, 33, 34, 35] consists in identifying particular subsets of the network, called $k$-cores, each one obtained by recursively removing all the vertices of degree smaller than $k$, until the degree of all remaining vertices is larger than or equal to $k$ (nodes pruned at step $k$ have shell index $k$ and form the $k$-shell). This decomposition therefore provides a probe to study the hierarchical properties of large scale networks, focusing on the network’s regions of increasing centrality and connectedness properties. It has been recently used as a basis for the visualization of large networks, in particular for AS maps [37, 36, 38]. Recent works using the $k$-core analysis have also focused on the analysis of the Internet maps obtained by the DIMES project [6], and an approach based on the $k$-core decomposition has even
Figure 6: Original and sampled clustering spectra for a heterogeneous network with large clustering (BA-DM, top) and for a BA network (with small clustering, bottom). Both networks have size $N = 10000$. The shape of the clustering spectra are globally preserved with both SP and RDM. For clustered networks, RDM yields better similarity between sampled and original networks. For the BA network, the sampling yields larger clustering coefficients, but does not change the order of magnitude.

been put forward to provide a conceptual and structural model of the Internet, the so-called Medusa model for the Internet [40, 39]. It seems therefore crucial to understand how the sampling of a network affects the properties of this decomposition. In [41], a first investigation in the case of SP sampling has revealed that the qualitative features of the decomposition are preserved by sampling, as soon as the sampling effort is reasonable. In particular, the very different fingerprints of such decomposition obtained with homogeneous or heterogeneous networks are preserved.

In Figure 7, we show the sizes of the $k$-shells as a function of $k$ for a random scale-free network and for the corresponding sampled networks obtained with various numbers of sources and targets. While the global shape of the curve is preserved by sampling, two interesting effects emerge. First, at a given sampling effort $\epsilon$, a larger number of sources leads to a sampled network with a larger maximal shell index, although the total number of discovered links and nodes decreases. This is similar to what happens e.g. for the estimation of the average degree; in the case of the $k$-core decomposition, a larger $N_S + N_T$ allows to discover better a smaller part of the network (note that in the extreme case of $N_S = 1$ or $N_T = 1$, the sampled graph is a tree with a maximal shell index of 1). The second observation concerns the comparison of SP and RDM sampling: even at small number of sources, the $k$-core decomposition of the sampled network is much closer to the original one for RDM, and in particular the maximal shell index is quite close to the real one. This shows that the quantitative bias observed in [41], which tends for SP sampling to underestimate the shell index of nodes, is partially corrected by a more realistic model for the sampling process: the structure of the $k$-core decomposition of Internet measured recently [39] is therefore most probably reflecting quite accurately the reality.

5 Conclusions

In this report, we have investigated in details how various ways of modeling the traceroute-like sampling of networks affect the comparison between sampled and original network. While most studies performed until now had focused on the assumption that traceroute probes go along the shortest paths (SP) from sources to destinations, we have considered a more realistic model in which all the deviations from

As shown in [36, 41], the $k$-core decomposition of the BA network is trivial, therefore here we use another type of scale-free network.
shortest-path routing, which come from complicate routing policies, congestion, etc, are effectively encoded into a small amount of random deviations. This random deviation model (RDM) has indeed been shown to reproduce accurately many properties of the real paths sampled by traceroute [24].

The comparison between SP and RDM sampling shows that the global qualitative picture obtained in [29, 30] is essentially unchanged. The fundamental reason is that, due to the introduction of random deviations, the probability for nodes and links to be discovered depends on a centrality measure that interpolates between the shortest path betweenness centrality and the random walk betweenness centrality. Since both quantities are strongly correlated, and as well correlated with the degree, the tails of the degree distributions are particularly well sampled, so that homogeneous and heterogeneous networks are easily discriminated, even at low levels of sampling. In fact, RDM allows for an even better discovery of degree distribution tails than SP, because the random deviations allow to discover links that are the support of only few shortest paths. At a quantitative level, RDM also allows to systematically discover larger parts of the network than RD, and in particular to obtain larger clustering coefficients, since the random deviations may cause the probes to discover triangles that would not be seen by shortest paths (see figure 5). This effect is particularly strong for triangles formed by small degree nodes, so that the clustering spectrum is strongly increased in this model. A rationalization of the values obtained in Internet sampling is therefore obtained, since such large values could hardly be met by SP sampling. Finally, the sampled $k$-core decomposition is also quantitatively closer to the real one with RDM, especially when the number of sources is large, bringing some more confidence that the values and characteristics recently measured in distributed projects such as DIMES [6, 39] are fairly representative of the real Internet.

In summary, while previous approximate analytical and numerical modeling of traceroute-like sampling of networks had already given a lot of evidence that the complex features observed in Internet maps were not an artifact of the mapping strategies, this work shows that more realistic models of the real traceroute tool lead to similar results. In fact, not only these conclusions are not affected, but they are strengthened since the sampled network is then even quantitatively closer to the real one.

References

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

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


