Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Computer Science 355 (2006) 6–24
Theoretical Computer Science
www.elsevier.com/locate/tcs

Exploring networks with traceroute-like probes: Theory and simulations

Luca Dall'Asta[a], Ignacio Alvarez-Hamelin[a], Alain Barrat[a,*], Alexei Vázquez[b], Alessandro Vespignani[a,c]

[a] *Laboratoire de Physique Théorique (UMR 8627 du CNRS), Bâtiment 210, Université de Paris-Sud, 91405 Orsay, Cedex, France*
[b] *Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN 46556, USA*
[c] *School of Informatics and Department of Physics, Indiana University, Bloomington, IN 47408, USA*

Abstract

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 destination, has been argued to introduce uncontrolled sampling biases that might produce statistical properties of the sampled graph which sharply differ from the original ones. In this paper, we explore these biases and provide a statistical analysis of their origin. We derive an analytical approximation for the probability of edge and vertex detection that exploits the role of the number of sources and targets and allows us to relate the global topological properties of the underlying network with the statistical accuracy of the sampled graph. In particular, we find that the edge and vertex detection probability depends on the *betweenness centrality* of each element. This allows us to show that shortest path routed sampling provides a better characterization of underlying graphs with broad distributions of connectivity. We complement the analytical discussion with a throughout numerical investigation of simulated mapping strategies in network models with different topologies. We show that sampled graphs provide a fair qualitative characterization of the statistical properties of the original networks in a fair range of different strategies and exploration parameters. Moreover, we characterize the level of redundancy and completeness of the exploration process as a function of the topological properties of the network. Finally, we study numerically how the fraction of vertices and edges discovered in the sampled graph depends on the particular deployements of probing sources. The results might hint the steps toward more efficient mapping strategies.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Traceroute; Internet exploration; Topology inference

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

* Corresponding author.
E-mail address: alain.barrat@th.u-psud.fr (A. Barrat).

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2005.12.009

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, 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. 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 shortest 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.

By using this strategy, a number of research groups have generated maps of the Internet [30,29,31,28,20], 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 $2 \leqslant \gamma \leqslant 2.5$ [14]. Evidence for the heavy-tailed behavior of the degree distribution has been collected in several other studies at the router and AS level [17,6,8,26,9] and have generated a large activity in the field of network modeling and characterization [23,21,11,2,26,3].

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 [7], typical mapping projects are run from relatively small sets of sources whose combined views are missing a considerable number of edges and vertices [9,33]. 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 [22,10,27]. Very interestingly, it has been shown that apparent degree distributions with heavy tails may be observed even from homogeneous topologies such as in the classic Erdös–Rényi (ER) graph model [22,10,1]. These studies thus point 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 this work we tackle this problem by performing a mean-field statistical analysis and extensive numerical study of shortest path routed sampling, considered as the first approximation to `traceroute`-sampling (see Section 3), in different networks models. We derive in Section 4 an approximate expression for the probability of edges and vertices to be detected that exploits the dependence upon the number of sources, targets and the topological properties of the networks. The expression shows the dependency of the efficiency of the mapping process upon the number of sources, targets and the topological properties of the network. Moreover, the analytical study provides a general understanding of which kind of topologies yields the most accurate sampling. In particular, we show that the map accuracy depends on the underlying network *betweenness centrality* distribution; the heavier the tail the higher the statistical accuracy of the sampled graph.

We substantiate our analytical finding with a throughout exploration of maps obtained varying the number of source–target pairs on networks models with different topological properties. In particular, we consider networks with degree distribution with poissonian, Weibull and power-law behavior. According to the theoretical analysis, both the total number of probes deployed and the topological properties seem to play a primary role in the understanding of the level of the efficiency reached by the mapping process. As a measure of the efficiency of the mapping in different network topologies, we study the fractions of discovered vertices and edges as a function of the degree (Section 5), stressing the agreement with the theoretical predictions. Other interesting quantities such as transit frequency and traffic entropy, are introduced in the study of the discovery process, with the aim of providing a complete framework for the study of sampling redundancy (Section 6). Furthermore, we focus on the study of the degree distributions obtained in the sampled graph and their resemblance to the original ones (see Section 7). Our results show that single source mapping processes face serious limitations in that also the targeting of the whole network results in a very partial discovery of its connectivity. On the contrary, the use of multiple sources promptly leads to obtained maps fairly consistent with the original sample, where the statistical degree distributions are qualitatively discriminated also at relatively low values

of target density. A detailed discussion of the behavior of the degree distribution as a function of targets and sources is provided for sampled graphs with different topologies and compared with the insight obtained by analytical means.

In Section 8, we also inspect quantitatively the portion of discovered network in different mapping strategies for the deployment of sources that however impose the same density of probes to the network; i.e. having the same probing load. We find the presence of a region of low efficiency (less vertices and edges discovered) depending on the relative proportion of sources and targets. This low efficiency region however corresponds to the optimal estimation of the network average degree. This finding calls for a "trade-off" between the accuracy in the observation of different quantities and hints to possible optimization procedures in the `traceroute`-driven mapping of large networks.

2. Related work

In this section, we shortly review some recent works devoted to the sampling of graphs by shortest path probing procedures. Lakhina et al. [22] have shown that biases can seriously affect the estimation of degree distributions. In particular, power-law like distributions can be observed for subgraphs of ER random graphs when the subgraph is the product of a `traceroute` exploration with relatively few sources and destinations. They discuss the origin of these biases and the effect of the distance between source and target in the mapping process. In a recent work [10], Clauset and Moore have given analytical foundations to the numerical work of Lakhina et al. [22] (see also [1]). They have modeled the single source probing to all possible destinations using differential equations. For an Erdös–Renyi random graph with average degree $\overline{k}$, they have found that the connectivity distribution of the obtained spanning tree displays a power-law behavior k^{-1}, with an exponential cut-off setting in at a characteristic degree $k_c \sim \overline{k}$.

In a slightly different context, Petermann and De Los Rios have studied a `traceroute`-like procedure on various examples of scale-free graphs [27], showing that, in the case of a single source, power-law distributions with under-estimated exponents are obtained. Analytical estimates of the measured exponents as a function of the true ones were also derived. Finally, a recent preprint by Guillaume and Latapy [18] reports about the shortest-paths explorations of synthetic graphs, focusing on the comparison between properties of the resulting sampled graph with those of the original network. The proportion of discovered vertices and edges in the graph as a function of the number of sources and targets gives also hints for an optimization of the exploration process.

All these pieces of work make clear the relevance of determining up to which extent the topological properties observed in sampled graphs are representative of that of the real networks.

3. Network models and `traceroute`-like processes

In a typical `traceroute` study, a set of active sources deployed in the network sends `traceroute` probes to a set of destination vertices. Each probe collects information on all the vertices and edges traversed along the path connecting the source to the destination, allowing the discovery of the network [7]. By merging the information collected on each path it is then possible to reconstruct a partial map of the network (Fig. 1). More in detail, the edges and the vertices discovered by each probe will depend on the "path selection criterium" used to decide the path between a pair of vertices. 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 even considerably from the shortest path. Despite these local, often unpredictable path distortions or inflations, a reasonable first approximation of the route traversed by `traceroute`-like probes is the shortest path between the two vertices. This assumption, however, is not sufficient for a proper definition of a `traceroute` model in that equivalent shortest paths between two vertices may exist. In the presence of a degeneracy of shortest paths we must therefore specify the path selection criterium by providing a resolution algorithm for the selection of shortest paths.

For the sake of simplicity we can define three selection mechanisms defining different ideal-paths that may account for some of the features encountered in Internet discovery:

- Unique shortest path (USP) probe. In this case the shortest path route selected between a vertex i and the destination target T is always the same independently of the source S (the path being initially chosen at random among all the equivalent ones).

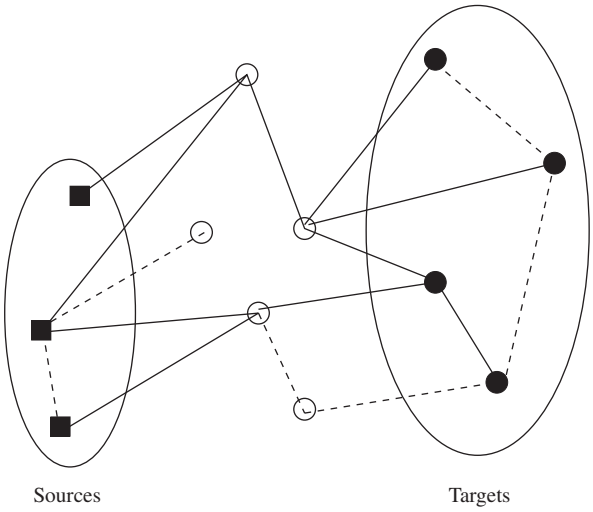

Fig. 1. Illustration of the `traceroute`-like procedure. Shortest paths between the set of sources and the set of destination targets are discovered (shown in full lines) while other edges are not found (dashed lines). Note that not all shortest paths are found since the "unique shortest path" procedure is used.

- Random shortest path (RSP) probe. The shortest path between any source–destination pair is chosen randomly among the set of equivalent shortest paths. This might mimic different peering agreements that make independent the paths among couples of vertices.
- All shortest paths (ASP) probe. The selection criterium discovers all the equivalent shortest paths between source–destination pairs. This might happen in the case of probing repeated in time (long time exploration), so that back-up paths and equivalent paths are discovered in different runs.

We will generically call $\mathcal{M}$-path the path found using one of these measurement or path selection mechanism. Actual `traceroute` probes contain a mixture of the three mechanisms defined above. We do not attempt, however, to account for all the subtleties that real studies encounters, i.e. IP routing, BGP policies, interface resolutions and many others. In fact, in the real mapping process, many effective heuristic strategies are commonly applied to improve the reliability and the performances of the sampling. For instance, the interface resolution is well achieved by the *iffinder* algorithm proposed by Broido and Claffy [6]. However, we will see that the different path selection criteria (p.s.c.) have only little influence on the general picture emerging from our results. Moreover, the USP procedure clearly represents the worst case scenario since, among the three different methods, it yields the minimum number of discoveries. For this reason, if not otherwise specified, we will report the USP data to illustrate the general features of our synthetic exploration. The interest of this analysis resides properly in the choice of working in the most pessimistic case, being aware that path inflations should actually provide a more pervasive sampling of the real network.

More formally, the experimental setup for our simulated `traceroute` mapping is the following. Let $G = (V, E)$ be a sparse undirected graph with vertices (vertices) $V = \{1, 2, \ldots, N\}$ and edges (links) E. Then let us define the sets of vertices $\mathcal{S} = \{i_1, i_2, \ldots, i_{N_S}\}$ and $\mathcal{T} = \{j_1, j_2, \ldots, j_{N_T}\}$ specifying the random placement of N_S sources and N_T destination targets. For each ensemble of source–target pairs $\Omega = \{\mathcal{S}, \mathcal{T}\}$, we compute with our p.s.c. the paths connecting each source–target pair. The sampled graph $\mathcal{G} = (V^*, E^*)$ is defined as the set of vertices V^* (with $N^* = |V^*|$) and edges E^* induced by considering the union of all the $\mathcal{M}$-paths connecting the source–target pairs. The sampled graph is thus analogous to the maps obtained from real `traceroute` sampling of the Internet.

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. 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 ρ_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 ρ_T allows us to compare mapping processes on networks with different sizes by defining an intrinsic percentage

Table 1
Main characteristics of the graphs used in the numerical exploration

	ER	ER	RSF	Weibull
N	10^4	10^4	10^4	10^4
$\|E\|$	10^5	5×10^5	22,000	55,000
$\bar{k}$	20	100	4.4	11
$k_{\max}$	40	140	3500	2000

of targeted vertices. In many cases, as we will see in the next sections, an appropriate quantity representing the level of sampling of the networks is $\varepsilon = N_S N_T / N$, 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.

In the following, our aim is to evaluate to which extent the statistical properties of the sampled graph $\mathcal{G}$ depend on the parameters of our experimental setup and are representative of the properties of the underlying graph G. The analytical insights gained in Section 4 will be complemented by a numerical investigation of the `traceroute`-like exploration process on various graph models endowed with very well-defined topological properties, so as to give a clear result on which kind of topologies are related to good sampling performances and vice versa. Starting from this first investigation, further studies could deal with more realistic models as those created using Internet topology generators [21,23]. In particular, we will consider two main classes of graphs.

(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. The most widely known model for homogeneous graphs is given by the classical ER model [13]: in such random graphs $G_{N,p}$ of N vertices, each edge is present in E independently with probability p. The expected number of edges is therefore $|E| = pN(N-1)/2$. In order to have sparse graphs one thus needs to have p of order $1/N$, since the average degree is $p(N-1)$. Erdös–Rényi graphs are typical examples of homogeneous graphs, with degree distribution following a Poisson law. Since $G_{N,p}$ can consist of more than one connected component, we consider only the largest of these components.

(B) *Heterogeneous graphs* for which $P(k)$ is a broad distribution with heavy tail and large fluctuations, spanning various orders of magnitude. In the literature, different definitions of heavy-tailed like distributions exist. 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) $P(k) = (a/c)(k/c)^{a-1} \exp(-(k/c)^a)$. The scale-free distribution has a diverging second moment and therefore virtually unbounded fluctuations, limited only by eventual size-cut-off. The Weibull distribution is akin to power-law distributions truncated by an exponential cut-off which are often encountered in the analysis of scale-free systems in the real world. Indeed, a truncation of the power-law behavior is generally due to finite-size effects and other physical constraints. Both forms have been proposed as representing the topological properties of the Internet [6]. We have generated the corresponding random graphs by using the algorithm proposed by Molloy and Reed [24]: the vertices of the graph are assigned a fixed sequence of degrees $\{k_i\}$, $i = 1, \ldots, N$, chosen at random from the desired degree distribution $P(k)$, and with the additional constraint that the sum $\sum_i k_i$ must be even; then, the vertices are connected by $\sum_i k_i/2$ edges, respecting the assigned degrees and avoiding self- and multiple-connections. The parameters used are $a = 0.25$ and $c = 0.6$ for the Weibull distribution, and $\gamma = 2.3$ for the RSF case. The main properties of the various graphs are summarized in Table 1. In all numerical studies we have used networks of $N = 10^4$ vertices. It is noteworthy that the maximum value of the degree ($k_{\max}$) is of the same order as the average for homogeneous graphs, but much larger for heterogenous ones.

4. Mean-field theory of simulated mapping process

We begin our study by presenting a mean-field statistical analysis of the simulated `traceroute` mapping. Our aim is to provide a statistical estimate for the probability of edge and vertex detection as a function of N_S, N_T and the topology of the underlying graph.

Let us define the quantity $\sigma_{i,j}^{(l,m)}$ that takes the value 1 if the edge (i, j) belongs to the selected $\mathcal{M}$-path between vertices l and m, and 0 otherwise. For a given set of sources and targets $\Omega = \{\mathcal{S}, \mathcal{T}\}$, the indicator function that a given edge (i, j) will be discovered and belongs to the sampled graph is simply $\pi_{i,j} = 1$ if the edge (i, j) belongs to at least one of the $\mathcal{M}$-paths connecting the source–target pairs, and 0 otherwise. We can obtain an exact expression for $\pi_{i,j}$ by noting that $1 - \pi_{i,j}$ is 1 if and only if (i, j) does not belong to any of the paths between sources and targets, i.e. if and only if $\sigma_{i,j}^{(l,m)} = 0$ for all $(l, m) \in \Omega$. This leads to

$$\pi_{i,j} = 1 - \prod_{l \neq m} \left(1 - \sum_{s=1}^{N_S} \delta_{l,i_s} \sum_{t=1}^{N_T} \delta_{m,j_t} \sigma_{i,j}^{(l,m)} \right), \tag{1}$$

where $\delta_{i,j}$ is the Kronecker symbol and selects only vertices belonging to the set of sources or targets.

While the above exact formula does not lead us too far in the understanding of the discovery probabilities, it is interesting to look at the process on a statistical ground by studying the average over all possible realizations of the set $\Omega = \{\mathcal{S}, \mathcal{T}\}$. By definition we have that

$$\left\langle \sum_{t=1}^{N_T} \delta_{i,j_t} \right\rangle = \rho_T \quad \text{and} \quad \left\langle \sum_{s=1}^{N_S} \delta_{i,i_s} \right\rangle = \rho_S, \tag{2}$$

where $\langle \cdots \rangle$ identifies the average over all possible deployment of sources and targets Ω. These equalities simply state that each vertex i has, on average, a probability to be a source or a target that is proportional to their respective densities. In the following, we will make use of an uncorrelation assumption that yields an explicit approximation for the discovery probability. The assumption consists in neglecting correlations originated by the position of sources and targets on the discovery probability by different paths. While this assumption does not provide an exact treatment for the problem it generally conveys a qualitative understanding of the statistical properties of the system. In this approximation, the average discovery probability of an edge is

$$\langle \pi_{i,j} \rangle = 1 - \left\langle \prod_{l \neq m} \left(1 - \sum_{s=1}^{N_S} \delta_{l,i_s} \sum_{t=1}^{N_T} \delta_{m,j_t} \sigma_{i,j}^{(l,m)} \right) \right\rangle \simeq 1 - \prod_{l \neq m} (1 - \rho_T \rho_S \langle \sigma_{i,j}^{(l,m)} \rangle), \tag{3}$$

where in the last term we take advantage of neglecting correlations by replacing the average of the product of variables with the product of the averages and using Eq. (2). This expression simply states that each possible source–target pair weights in the average with the product of the probability that the end vertices are a source and a target; the discovery probability is thus obtained by considering the edge in an average effective media (*mean-field*) of sources and targets homogeneously distributed in the network. This approach is indeed akin to mean-field methods customarily used in the study of many particle systems where each particle is considered in an effective average medium defined by the uncorrelated averages of quantities. The realization average of $\langle \sigma_{i,j}^{(l,m)} \rangle$ is very simple in the uncorrelated picture, depending only of the kind of the probing model. In the case of the ASP probing, $\langle \sigma_{i,j}^{(l,m)} \rangle$ is just one if (i, j) belongs to one of the shortest paths between l and m, and 0 otherwise. In the case of the USP and the RSP, on the contrary, only one path among all the equivalent ones is chosen. If we denote by $\sigma^{(l,m)}$ the number of shortest paths between vertices l and m, and by $x_{i,j}^{(l,m)}$ the number of these paths passing through the edge (i, j), the probability that the `traceroute` model chooses a path going through the edge (i, j) between l and m is $\langle \sigma_{i,j}^{(l,m)} \rangle = x_{i,j}^{(l,m)} / \sigma^{(l,m)}$.

The standard situation we consider is the one in which $\rho_T \rho_S \ll 1$ and since $\langle \sigma_{i,j}^{(l,m)} \rangle \leqslant 1$, we have

$$\prod_{l \neq m} (1 - \rho_T \rho_S \langle \sigma_{i,j}^{(l,m)} \rangle) \simeq \prod_{l \neq m} \exp(-\rho_T \rho_S \langle \sigma_{i,j}^{(l,m)} \rangle), \tag{4}$$

that inserted in Eq. (3) yields

$$\langle \pi_{i,j} \rangle \simeq 1 - \prod_{l \neq m} (\exp(-\rho_T \rho_S \langle \sigma_{i,j}^{(l,m)} \rangle)) = 1 - \exp(-\rho_T \rho_S b_{ij}), \tag{5}$$

where $b_{ij} = \sum_{l \neq m} \langle \sigma_{i,j}^{(l,m)} \rangle$. In the case of the USP and RSP probing, the quantity b_{ij} is by definition the edge *betweenness centrality* $\sum_{l \neq m} x_{i,j}^{(l,m)} / \sigma^{(l,m)}$ [15,5], sometimes also refereed to as "load" [16] (in the case of ASP

probing, it is a closely related quantity). Indeed the vertex or edge betweenness is defined as the total number of shortest paths among pairs of vertices in the network that pass through a vertex or an edge, respectively. If there are multiple shortest paths between a pair of vertices, the path contributes to the betweenness with the corresponding relative weight. The *betweenness* gives a measure of the amount of all-to-all traffic that goes through an edge or vertex, if the shortest path is used as the metric defining the optimal path between pairs of vertices, and it can be considered as a non-local measure of the *centrality* of an edge or vertex in the graph.

The edge betweenness assumes values between 2 and $N(N-1)$ and the discovery probability of the edge will therefore depend strongly on its betweenness. In particular, for edges with minimum betweenness $b_{ij} = 2$ we have $\langle \pi_{i,j} \rangle \simeq 2\rho_{\mathrm{T}}\rho_{\mathrm{S}}$, that recovers the probability that the two end vertices of the edge are chosen as source and target. This implies that if the densities of sources and targets are small but finite in the limit of very large N, all the edges in the underlying graph have an appreciable probability to be discovered. Moreover, for edges with high betweenness the discovery probability approaches one. A fair sampling of the network is thus expected.

In most realistic samplings, however, we face a very different situation. While it is reasonable to consider ρ_{T} a small but finite value, the number of sources is not extensive ($N_{\mathrm{S}} \sim \mathcal{O}(1)$) and their density tends to zero as N^{-1}. In this case it is more convenient to express the edge discovery probability as

$$\langle \pi_{i,j} \rangle \simeq 1 - \exp\left(-\varepsilon \widetilde{b_{ij}}\right), \tag{6}$$

where $\varepsilon = \rho_{\mathrm{T}} N_{\mathrm{S}}$ is the density of probes imposed to the system and the rescaled betweenness $\widetilde{b_{ij}} = N^{-1} b_{ij}$ is now limited in the interval $[2N^{-1}, N-1]$. In the limit of large networks $N \to \infty$ it is clear that edges with low betweenness have $\langle \pi_{i,j} \rangle \sim \mathcal{O}(N^{-1})$, for any finite value of ε. This readily implies that in real situations the discovery process is generally not complete, a large part of low betweenness edges being not discovered, and that the network sampling is made progressively more accurate by increasing the density of probes ε.

A similar analysis can be performed for the discovery probability π_i of a vertex i. For each source–target set Ω we have that

$$\pi_i = 1 - \left(1 - \sum_{s=1}^{N_{\mathrm{S}}} \delta_{i,i_s} - \sum_{t=1}^{N_{\mathrm{T}}} \delta_{i,j_t}\right) \prod_{l \neq m \neq i} \left(1 - \sum_{s=1}^{N_{\mathrm{S}}} \delta_{l,i_s} \sum_{t=1}^{N_{\mathrm{T}}} \delta_{m,j_t} \sigma_i^{(l,m)}\right). \tag{7}$$

where $\sigma_i^{(l,m)} = 1$ if the vertex i belongs to the $\mathcal{M}$-path between vertices l and m, and 0 otherwise. This time it has been considered that each vertex is discovered with probability one also if it is in the set of sources and targets. The second term on the right-hand side therefore expresses the fact that the vertex i does not belong to the set of sources and targets and it is not discovered by any $\mathcal{M}$-path between source–target pairs. By using the same *mean-field* approximation as previously, the average vertex discovery probability reads as

$$\langle \pi_i \rangle \simeq 1 - (1 - \rho_{\mathrm{S}} - \rho_{\mathrm{T}}) \prod_{l \neq m \neq i} \left(1 - \rho_{\mathrm{T}}\rho_{\mathrm{S}} \langle \sigma_i^{(l,m)} \rangle\right). \tag{8}$$

As for the case of the edge discovery probability, the average considers all possible source–target pairs weighted with probability $\rho_{\mathrm{T}}\rho_{\mathrm{S}}$. In the ASP model, the average $\langle \sigma_i^{(l,m)} \rangle$ is 1 if i belongs to one of the shortest paths between l and m, and 0 otherwise. For the USP and RSP models, $\langle \sigma_i^{(l,m)} \rangle = x_i^{(l,m)} / \sigma^{(l,m)}$ where $x_i^{(l,m)}$ is the number of shortest paths between l and m going through i. If $\rho_{\mathrm{T}}\rho_{\mathrm{S}} \ll 1$, by using the same approximations used for Eq. (5) we obtain

$$\langle \pi_i \rangle \simeq 1 - (1 - \rho_{\mathrm{S}} - \rho_{\mathrm{T}}) \exp\left(-\rho_{\mathrm{T}}\rho_{\mathrm{S}} b_i\right), \tag{9}$$

where $b_i = \sum_{l \neq m \neq i} \langle \sigma_i^{(l,m)} \rangle$. For the USP and RSP cases, $b_i = \sum_{l \neq m \neq i} x_i^{(l,m)} / \sigma^{(l,m)}$ is the vertex betweenness centrality, that is limited in the interval $[0, N(N-1)]$ [15,5,16]. The betweenness value $b_i = 0$ holds for the leafs of the graph, i.e. vertices with a single edge, for which we recover $\langle \pi_i \rangle \simeq \rho_{\mathrm{S}} + \rho_{\mathrm{T}}$. Indeed, this kind of vertices are dangling ends discovered only if they are either a source or target themselves.

As discussed before, the most usual setup corresponds to a density $\rho_{\mathrm{S}} \sim \mathcal{O}(N^{-1})$ and in the large N limit we can conveniently write

$$\langle \pi_i \rangle \simeq 1 - (1 - \rho_{\mathrm{T}}) \exp\left(-\varepsilon \widetilde{b_i}\right), \tag{10}$$

where we have neglected terms of order $\mathcal{O}(N^{-1})$ and the rescaled betweenness $\widetilde{b_i} = N^{-1}b_i$ is now defined in the interval $[0, N-1]$. This expression points out that the probability of vertex discovery is favored by the deployment of a finite density of targets that defines its lower bound.

We can also provide a simple approximation for the effective average degree $\langle k_i^* \rangle$ of the vertex i discovered by our sampling process. Each edge departing from the vertex will contribute proportionally to its discovery probability, yielding

$$\langle k_i^* \rangle = \sum_j (1 - \exp(-\varepsilon \widetilde{b_{ij}})) \simeq \varepsilon \sum_j \widetilde{b_{ij}}. \tag{11}$$

The final expression is obtained for edges with $\varepsilon \widetilde{b_{ij}} \ll 1$. Since the sum over all neighbors of the edge betweenness is simply related to the vertex betweenness as $\sum_j b_{ij} = 2(b_i + N - 1)$, where the factor 2 considers that each vertex path traverses two edges and the term $N-1$ accounts for all the edge paths for which the vertex is an endpoint, this finally yields

$$\langle k_i^* \rangle \simeq 2\varepsilon + 2\varepsilon \widetilde{b_i}. \tag{12}$$

The present analysis shows that the measured quantities and statistical properties of the sampled graph strongly depend on the parameters of the experimental setup and the topology of the underlying graph. The latter dependence is exploited by the key role played by edge and vertex betweenness in the expressions characterizing the graph discovery. The betweenness is a non-local topological quantity whose properties change considerably depending on the kind of graph considered. This allows an intuitive understanding of the fact that graphs with diverse topological properties deliver different answer to sampling experiments.

5. Efficiency of the mapping process

The analytical findings of the previous section may be tested and used as guidance in the numerical analysis of simulated mapping experiments of network models. In particular we will consider the graph topologies defined in Section 3. Let us first consider the case of homogeneous graphs (ER model): as shown in Fig. 2, the vertex and edge betweennesses are homogeneous quantities and their distributions are peaked around their average values $\overline{b}$ and $\overline{b_e}$, respectively, spanning only a small range of variations. These values can thus be considered as typical values. We can thus use Eqs. (6) and (10) to estimate the order of magnitude of probes that allows a fair sampling of the graph. Indeed, both $\langle \pi_{i,j} \rangle$ and $\langle \pi_i \rangle$ tend to 1 if $\varepsilon \gg \max\left[\overline{b}^{-1}, \overline{b_e}^{-1}\right]$. In this limit all edges and vertices will have probability to be discovered very close to one. At lower value of ε, obtained by varying ρ_{T} and N_{S}, the underlying graph is only partially discovered. Fig. 3 shows the behavior of the fraction N_k^*/N_k of discovered vertices of degree k, where N_k is the total number of vertices of degree k in the underlying graph, and the fraction of discovered edges $\langle k^* \rangle / k$ in vertices of degree k. N_k^*/N_k naturally increases with the density of targets and sources, and it is slightly increasing with k. The latter behavior can be easily understood by noticing that vertices with larger degree have on average a larger betweenness (see Fig. 2). On the other hand, the range of variation of k in homogeneous graphs is very narrow and only a large level of probing may guarantee very large discovery probabilities. Similarly the behavior of the effective discovered degree can be understood by looking at Eq. (12). Indeed the initial decrease of $\langle k^* \rangle / k$ is finally compensated by the increase of $\overline{b(k)}$.

The situation is different in graphs with heavy-tailed connectivity distributions (RSF and WEI models), with an appreciable fraction of vertices and edges with very high betweenness [4]. In particular, in scale-free graphs the site betweenness is related to the vertices degree as $\overline{b(k)} \sim k^{\beta}$, where β is an exponent depending on the model [4]. Since in heavy-tailed degree distributions the allowed degree is varying over several orders of magnitude, the same occurs for the betweenness values, as shown in Fig. 2, and the tail of the distribution is broader the broader the connectivity distribution: larger values are consequently reached for the RSF case with $\gamma = 2.3$ than for the WEI case. In such a situation, even in the case of small ε, vertices whose betweenness is large enough ($b_i \varepsilon \gg 1$) have $\langle \pi_i \rangle \simeq 1$. Therefore, all vertices with degree $k \gg \varepsilon^{-1/\beta}$ will be detected with probability one. This is clearly visible in Fig. 3 where the discovery probability N_k^*/N_k of vertices with degree k saturates to one for large degree values. Consistently, the degree value at which the curve saturates decreases with increasing ε. A similar effect is appearing in the measurements

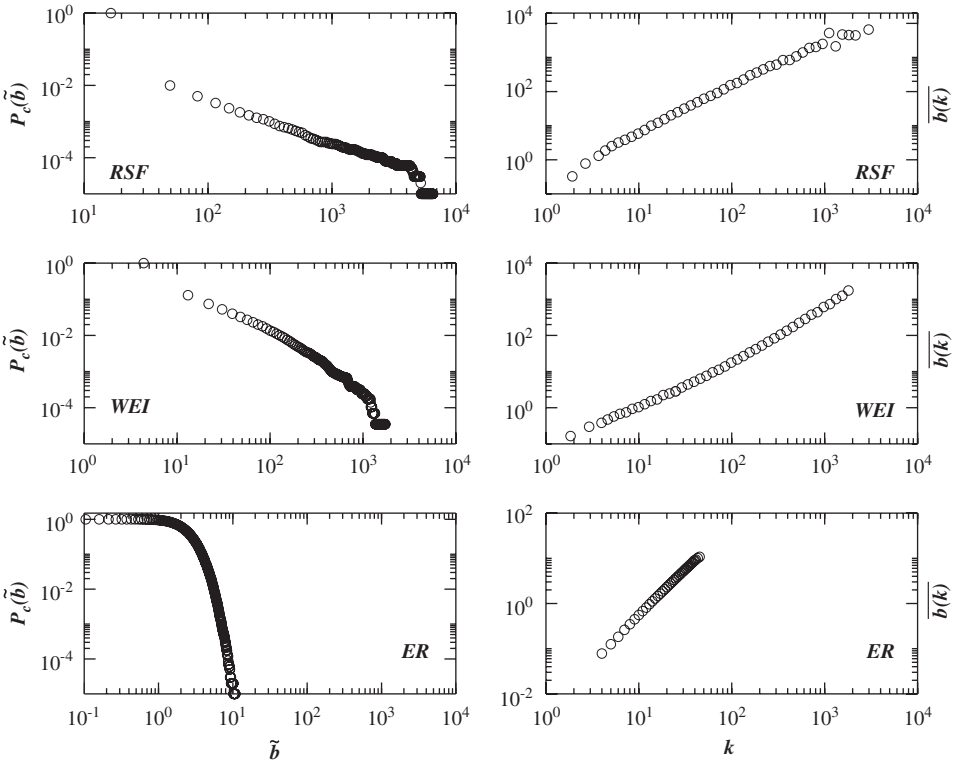

Fig. 2. Cumulative distribution of the rescaled vertex betweenness (left) and average behavior as a function of the connectivity (right) in the various graph models considered.

concerning $\langle k^* \rangle / k$. After an initial decay (Fig. 3) the effective discovered degree is increasing with the degree of the vertices. This qualitative feature is captured by Eq. (12) that gives $\langle k^* \rangle / k \simeq \varepsilon k^{-1}(1 + \overline{b(k)})$. At large k the term $k^{-1}\overline{b(k)} \sim k^{\beta-1}$ takes over and the effective discovered degree approaches the real degree k. Moreover, it appears clearly that the broader the distribution of betweennesses or connectivities, the better the sampling obtained.

6. Redundancy and dissymmetry in the mapping process

In this section, we introduce tools suitable to estimate how `traceroute`-like procedures discover the vertices and the edges of the unknown underlying network. The most common biases affecting the mapping process concern the miss of lateral connectivity, and the multiple sampling of central vertices (and edges), which may affect the efficiency of the whole process. While the first problem might be solved by an optimization in the deployment of probes, actually relying on a criterion of decentralization of sources and targets, multiple sampling can be studied through some general concepts like the *redundancy* and *dissymmetry* of the discovery process.

6.1. Redundancy

Let us define the edge redundancy $r_e(i, j)$ of an edge (i, j) in a `traceroute`-sampling as the number of probes passing through the edge (i, j). Using the notations of Section 4, this quantity is written for a given set of probes and targets as

$$r_e(i, j) = \sum_{l \neq m} \left(\sum_{s=1}^{N_S} \delta_{l,i_s} \sum_{t=1}^{N_T} \delta_{m,i_t} \sigma_{i,j}^{(l,m)} \right). \tag{13}$$

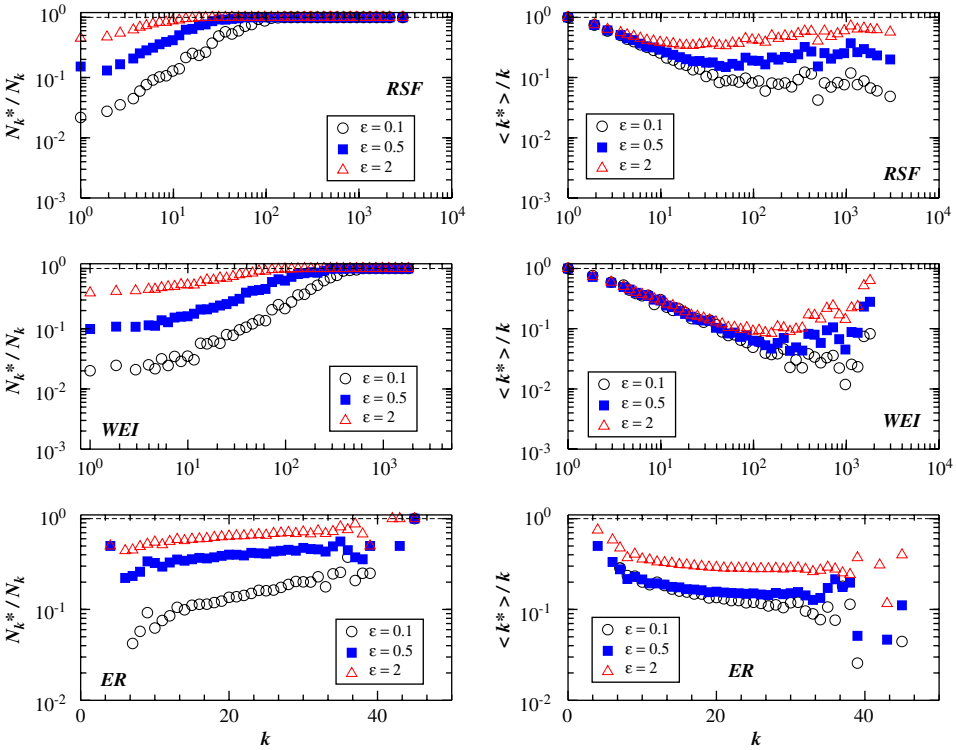

Fig. 3. Frequency N_k^*/N_k of detecting a vertex of degree k (left) and proportion of discovered edges $\langle k^* \rangle / k$ (right) as a function of the degree in the RSF, WEI, and ER graph models. The exploration setup considers $N_S = 2$ and increasing probing level ε obtained by progressively higher density of targets ρ_T. The axis of ordinates is in log scale to allow a finer resolution.

Averaging over all possible realizations and assuming the uncorrelation hypothesis, we obtain

$$\langle r_e(i, j) \rangle \simeq \sum_{l \neq m} \rho_T \rho_S \langle \sigma_{i,j}^{(l,m)} \rangle = \rho_T \rho_S b_{ij}. \tag{14}$$

This result implies that the average redundancy of an edge is related to the density of sources and targets, but also to the edge betweenness. For example, an edge of minimum betweenness $b_{ij} = 2$ can be discovered at most twice in the extreme limit of an all-to-all probing. On the contrary, a very central edge of betweenness b_{ij} close to the maximum $N(N-1)$, would be discovered with a redundancy close to $(N-1)$ by a `traceroute`-probing from a single source to all the possible destinations.

Similarly, the redundancy $r_n(i)$ of a vertex i, intended as the number of times the probes cross the vertex i, can be obtained:

$$r_n(i) = \sum_{l \neq m} \sigma_i^{(l,m)} \sum_{s=1}^{N_S} \delta_{l,i_s} \sum_{t=1}^{N_T} \delta_{m,i_t}. \tag{15}$$

After separating the cases $l = i$ and $m = i$ in the sum, the averaging over the positions of sources and targets yields in the mean-field approximation:

$$\langle r_n(i) \rangle = \sum_{l \neq m \neq i} \rho_S \rho_T \langle \sigma_i^{(l,m)} \rangle + 2\rho_S \rho_T N \simeq 2\varepsilon + \rho_S \rho_T b_i. \tag{16}$$

In this case, a term related to the number of traceroute probes ε appears, showing that a part of the mapping effort unavoidably ends up in generating vertex detection redundancy.

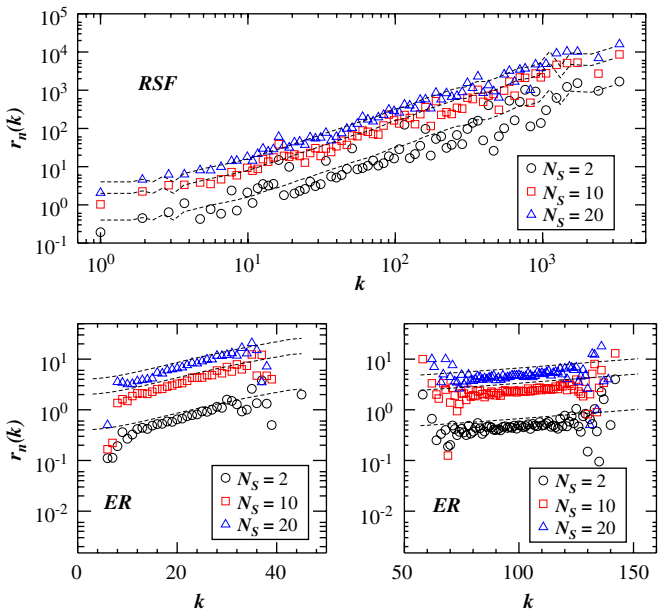

Fig. 4. Average vertex redundancy as a function of the degree k for RSF (top) and ER (bottom) model ($N = 10^4$). For the ER model, two blocks of data are plotted, for $\overline{k} = 20$ (left) and for $\overline{k} = 100$ (right). The target density is fixed ($\rho_T = 0.1$), and $N_S = 2$ (circles), 10 (squares), 20 (triangles). The dashed lines represent the analytical prediction $2\varepsilon + \rho_S \rho_T \overline{b(k)}$ in perfect agreement with the simulations.

In Fig. 4 we report the behavior of the average vertex redundancy as a function of the degree k for both homogeneous (ER) and heterogeneous (RSF) graphs. For both models, the behaviors are in good agreement with the mean-field prediction, showing the tight relation between redundancy and betweenness centrality.

In the case of heavy-tailed underlying networks, the vertex redundancy typically grows as a power-law of the degree, while the values for random graphs vary on a smaller scale. This behavior points out that the intrinsic hierarchical structure of scale-free networks plays a fundamental role even in the process of path routing, resulting in a huge number of probes iteratively passing through the same set of few hubs. On the other hand, for homogeneous graphs the total number of vertex discoveries is quite uniformly distributed on the whole range of connectivity, independently of the relative importance of the vertices.

6.2. Dissymmetry: participation ratio

The high rate of redundancy intrinsic to the exploration process, however, does not imply that the local topology close to a vertex is well discovered: preferential paths could indeed carry most of the probing effort leading to just a partial discovery of the vertex connections. This amounts to a dissymmetry of the exploration process that probes some edges much more than others, eventually ignoring some of those, in the neighborhood of a given vertex. Together with redundancy measures, let us consider the relative number of occurrences of a given edge (i, j) during the traceroute, with respect to the total occurrence for the edges in the neighborhood of i. For each discovered vertex i, we can thus define a set of frequencies $\{f_j^{(i)}\}_{j \in \mathcal{V}(i)}$ for the edges (i, j) of its neighborhood. In terms of redundancy the *edge frequency* $f_j^{(i)}$ is defined as

$$f_j^{(i)} = \frac{r_e(i, j)}{\sum_{j \in \mathcal{V}(i)} r_e(i, j)}, \quad 0 \leqslant f_j^{(i)} \leqslant 1 \quad \forall (i, j) \in E \tag{17}$$

and indicates the probability that any given probing path discovering the vertex i, is passing by the edge (i, j). The dissymmetry of the discovery of the neighborhood of a vertex may be quantified through the *participation ratio*

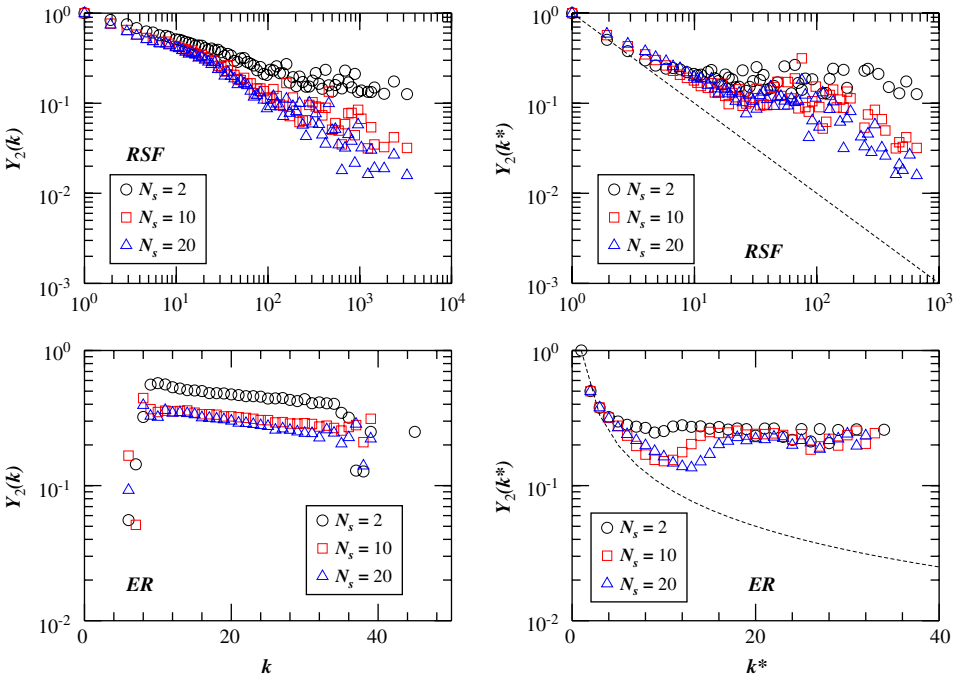

Fig. 5. Participation ratio as a function of real (k) and discovered (k^*) connectivity for RSF (top) and ER (bottom) models ($N = 10^4$). The target density is fixed ($\rho_T = 0.1$) and three value of N_S are presented: 2 (circles), 10 (squares), 20 (triangles). The dashed lines correspond to the $1/k^*$ bound.

of these frequencies:

$$Y_2(i) = \sum_{j \in \mathcal{V}(i)} (f_j^{(i)})^2. \tag{18}$$

If all the edge frequencies of i are of the same order $\sim 1/k_i^*$ (only discovered links give a finite contribution), the participation ratio should decrease as $1/k_i^*$ with increasing discovered connectivity k_i^*. Hence, in the limit of an optimally symmetric sampling, it should yield a strict power law behavior $Y_2(k^*) \sim k^{*-1}$. On the contrary, when only few links are preferred, for instance because more central in the shortest path routing, the sum is dominated by these terms, leading to a value closer to the upper bound 1. Numerical data for Y_2 as a function of the actual (k) and discovered (k^*) connectivity for different probing efforts, are displayed in Fig. 5. For heterogeneous graphs, the values of Y_2 tend towards the curve k^{*-1} for increasing ε. In both cases this behavior is better achieved at high degree values. The tendency of high degree vertices to be better sampled in a more symmetrical way is evident in the diagram for $Y_2(k)$, where a crossover at large degrees appears. On the contrary, in the homogeneous case (ER), the figures show a general high level of dissymmetry persistent at all degree values, only slightly dependent on the actual connectivity and the probing effort.

6.3. Dissymmetry: entropy measure

In order to provide an alternative and in some cases more accurate study of the dissymmetry of the exploration process, we introduce a more refined frequency, $f_{kj}^{(i)}$ defined as the number of probes passing through the *pair* $(k, i) - (i, j)$ of edges centered on the vertex i. This is the probability of a probe to traverse a couple of edges with respect to the total number of transits through any of the possible couples of edges in the neighborhood of i. This frequency takes fully into account the path traversing each vertex and the dissymmetry of the flow. By means of this frequency, we define an entropy measure providing supplementary evidence of the tight relation between local accuracy, dissymmetry of sampling and topological characterization of graphs. Indeed, a `traceroute` discovering vertices crossing a larger

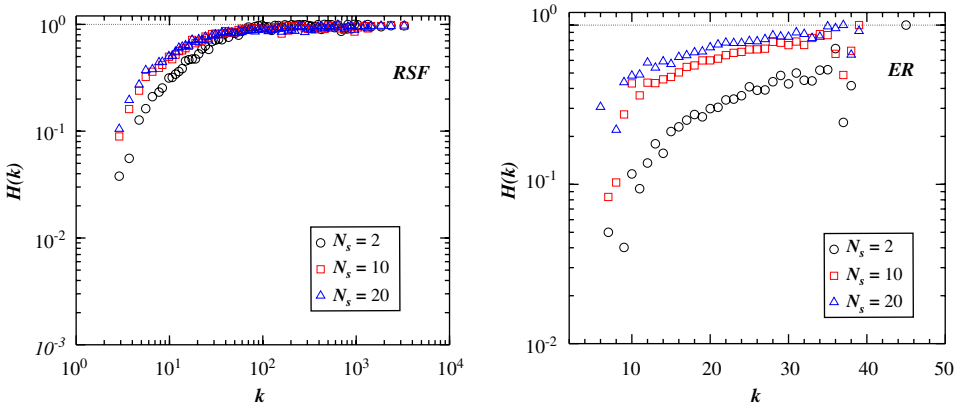

Fig. 6. Entropy vs. k: a saturation effect is clear at medium–high degree vertices for scale-free topologies (RSF), instead of a more regular increasing for homogeneous graphs (ER). In the figure there are different curves for $N_S = 2$ (circles), 10 (squares), 20 (triangles) and $\rho_T = 0.1$.

variety of their links, and with different paths, is expected to be more accurate (and likely efficient) than the one always selecting the same path. In the same spirit of the Shannon entropy, which is a good indicator of disorder, we define the *local traceroute entropy* of a vertex i by

$$h_i = -\frac{1}{\log(k_i^*(k_i^* - 1))} \sum_{k \neq j \in \mathcal{V}(i)} f_{kj}^{(i)} \log f_{kj}^{(i)}, \tag{19}$$

where $\log k_i^*$ is simply a normalization factor. This quantity is bounded in the interval $0 \leqslant h(i) \leqslant 1$. The case $h_i = 1$ is reached when all the frequencies of probes spanning the edge couples of the vertex are equal. The case $H \simeq 0$ corresponds to a dominating frequency in a specific edge couple. Also in this case it is possible to study the degree spectrum $H(k)$ of the entropy by measuring the average entropy on vertices with given degree k.

The numerical data of $H(k)$ for RSF and ER models and for different levels of probing are reported in Fig. 6. The values for ER are slightly increasing both for increasing degree k and number of sources N_S, with no qualitative difference in the behavior at low or high degree regions. On the other hand, the case of heterogeneous networks agrees with the previous observations. The curve for $H(k)$, indeed, shows a saturation phenomenon to values very close to the maximum 1 at large enough degree, indicating a very symmetric sampling of these vertices.

In summary the previous studies indicate that in the case of heterogeneous networks, the hubs and high betweenness vertices are in general sampled redundantly, however, obtaining a rather symmetrical discovery of their neighborhood. On the contrary, homogeneous networks do not allow the presence of hubs and vertices are suffering a less redundant sampling while showing a high dissymmetry of the local exploration process. This results might be useful in deciding source–target deployment strategies, by taking into account the underlying topology of the network.

7. Degree distribution measurements

A very important quantity in the study of the statistical accuracy of the sampled graph is the degree distribution. Fig. 7 shows the cumulative degree distribution $P_c(k^* > k)$ of the sampled graph defined by the ER model for increasing density of targets and sources. Sampled distributions are only approximating the genuine distribution, however, for $N_S \geqslant 2$ they are far from true heavy-tail distributions at any appreciable level of probing. Indeed, the distribution runs generally over a small range of degrees, with a cut-off that sets in at the average degree $\bar{k}$ of the underlying graph. In order to stretch the distribution range, homogeneous graphs with very large average degree $\bar{k}$ must be considered; however, other distinctive spurious effects appear in this case. In particular, since the best sampling occurs around the high degree values, the distributions develop peaks that show in the cumulative distribution as plateaus (see Fig. 8). Finally, in the case of RSP and ASP model, we observe that the obtained distributions are closer to the real one since they allow a larger number of discoveries.

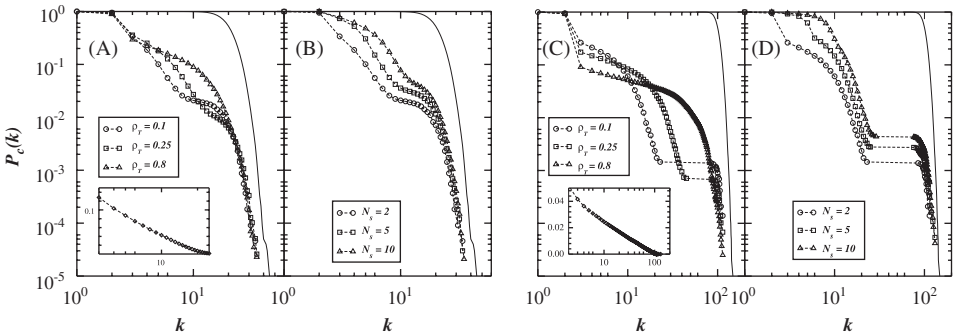

Fig. 7. Cumulative degree distribution of the sampled ER graph for USP probes. Panels (A) and (B) correspond to $\overline{k} = 20$, and (C) and (D) to $\overline{k} = 100$. Panels (A) and (C) show sampled distributions obtained with $N_S = 2$ and varying density target ρ_T. In the insets we report the peculiar case $N_S = 1$ that provides an apparent power-law behavior with exponent -1 at all values of ρ_T, with a cut-off depending on $\overline{k}$. The insets are in lin-log scale to show the logarithmic behavior of the corresponding cumulative distribution. Panels (B) and (D) correspond to $\rho_T = 0.1$ and varying number of sources N_S. The solid lines are the degree distributions of the underlying graph. For $\overline{k} = 100$, the sampled cumulative distributions display plateaus corresponding to peaks in the degree distributions, induced by the sampling process.

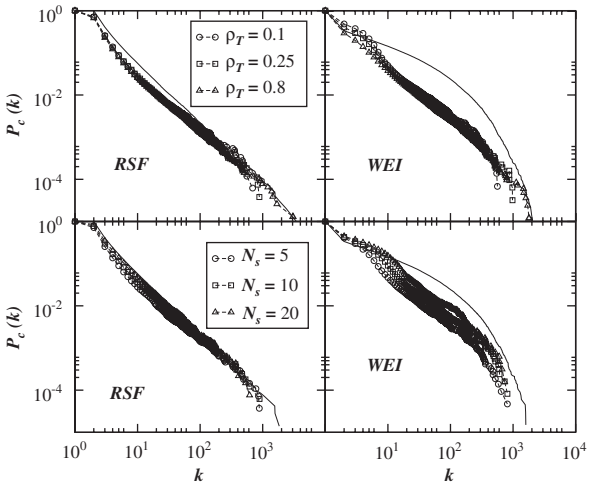

Fig. 8. Cumulative degree distributions of the sampled RSF and WEI graphs for USP probes. The top figures show sampled distributions obtained with $N_S = 5$ and varying density target ρ_T. The figures on the bottom correspond to $\rho_T = 0.25$ and varying number of sources N_S. The solid lines are the degree distributions of the underlying graph.

Only in the peculiar case of $N_S = 1$ an apparent scale-free behavior with slope -1 is observed for all target densities ρ_T, as analytically shown by Clauset and Moore [10,1]. Also in this case, the distribution cut-off is consistently determined by the average degree $\overline{k}$. It is worth noting that the experimental setup with a single source is a limit case corresponding to a highly asymmetric probing process; it is therefore badly, if at all, captured by our statistical analysis which assumes homogeneous deployment.

The present analysis shows that in order to obtain a sampled graph with apparent scale-free behavior on a degree range varying over n orders of magnitude we would need the very peculiar sampling of a homogeneous underlying graph with an average degree $\overline{k} \simeq 10^n$; a rather unrealistic situation in the Internet and many other information systems where $n \geqslant 2$.

In Section 5, we have shown clearly that, in heterogeneous graphs, vertices with high degree are efficiently sampled with an effective measured degree that is rather close to the real one. This means that the degree distribution tail is fairly well sampled while deviations should be expected at lower degree values. This is indeed what we observe in numerical

experiments on graphs with heavy-tailed distributions (see Fig. 8). Despite both RSF and WEI underlying graphs have a small average degree, the observed degree distribution spans more than two orders of magnitude. The distribution tail is fairly reproduced even at rather small values of ε. The data shows clearly that the low degree regime is instead undersampled. This undersampling can either yield an apparent change in the exponent of the degree distribution (as also noticed in [27] for single source experiments), or, if N_S is small, yield a power-law like distribution for an underlying Weibull distribution. Furthermore, as Fig. 8 shows, an increase in the number of sources starts to discriminate between scale-free and Weibull distributions by detecting a curvature in the second case even at small values $\rho_T = 0.25$. It is, however, fair to say that while the experiments clearly point out a broad and heavy-tailed distribution, the distinction between different types of heavy-tailed distribution needs an adequate level of probing.

In conclusion, graphs with heavy-tailed degree distribution allow a better qualitative representation of their statistical features in sampling experiments. Indeed, the most important properties of these graphs are related to the heavy-tail part of the statistical distributions that are indeed well discriminated by the `traceroute`-like exploration. On the other hand, the accurate identification of the distribution forms requires a fair level of sampling that it is not clear how to determine quantitatively in the case of an unknown underlying network. We will discuss the implications of these results in real Internet measurements in Section 9.

8. Optimization of mapping strategies

In the previous sections we have shown 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. Unfortunately, the mean-field approximation breaks down when we aim at a quantitative representation of the results. The neglected correlations are in fact very important for the precise estimate of the various quantities of interest. For this reason we performed an extensive set of numerical explorations aimed at a fine determination of the level of sampling achieved for different experimental setups.

In Fig. 9 we report the proportion of discovered edges in the numerical exploration of the graph models defined previously for increasing level of probing ε. The level of probing is increased either by raising the number of sources at fixed target density or by raising the target density at fixed number of sources. As expected, both strategies are progressively more efficient with increasing levels of probing. In heterogeneous graphs, it is also possible to see that when the number of sources is $N_S \sim \mathcal{O}(1)$ the increase of the number of targets achieves better sampling than increasing the deployed sources. On the other hand, it is easy to perceive that the shortest path route mapping is a symmetric process if we exchange sources with targets. This is confirmed by numerical experiments in which we use a very large number of sources and a number of targets $\rho_T \sim \mathcal{O}(1/N)$, where the trends are opposite: the increase of the number of sources achieves better sampling than increasing the deployed targets.

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. 10, 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 $g_\varepsilon(\rho_T) = g_\varepsilon(\varepsilon/N\rho_T)$. 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. 10 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

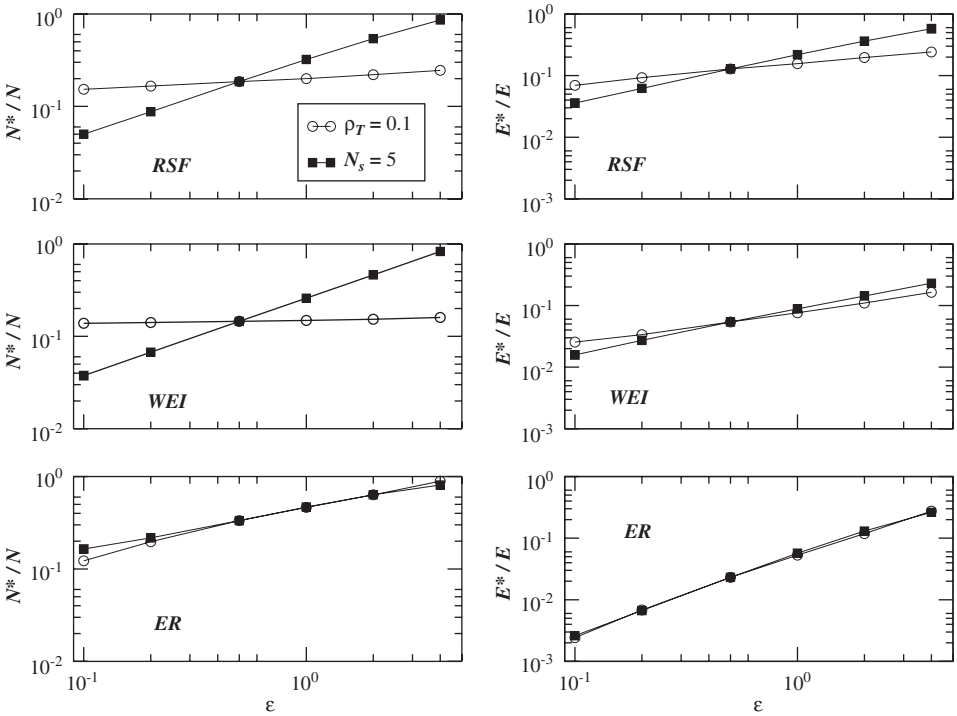

Fig. 9. Behavior of the fraction of discovered edges in explorations with increasing ε. For each underlying graph studied we report two curves corresponding to larger ε achieved by increasing the target density ρ_{T} at constant $N_{\mathrm{S}} = 5$ (squares) or the number of sources N_{S} at constant $\rho_{\mathrm{T}} = 0.1$ (circles).

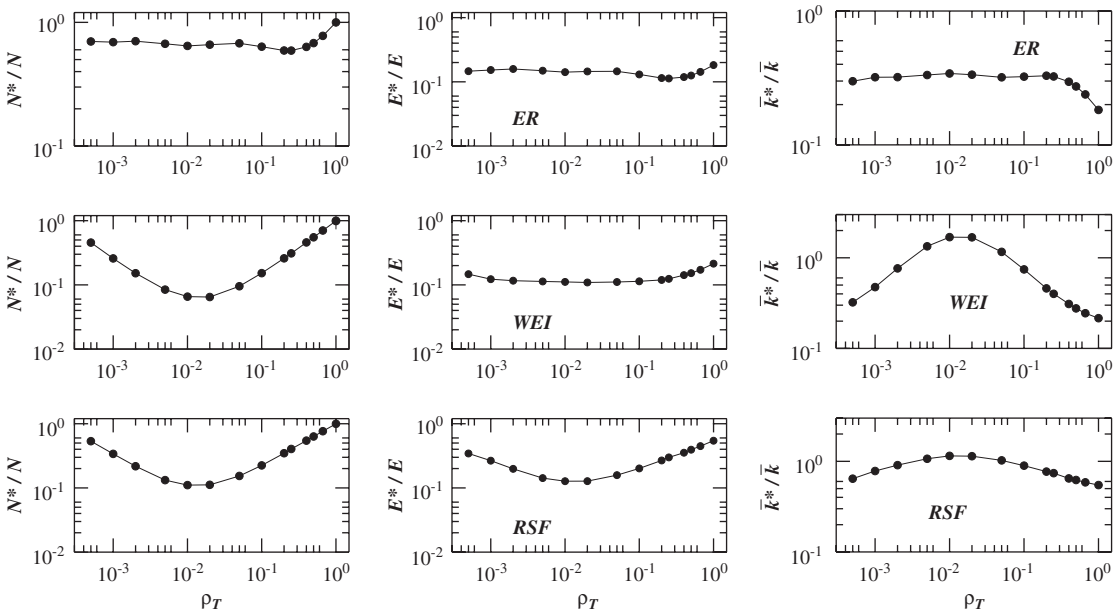

Fig. 10. Behavior as a function of ρ_{T} of the fraction of discovered edges and vertices in explorations with fixed ε (here $\varepsilon = 2$). Since $\varepsilon = \rho_{\mathrm{T}} N_{\mathrm{S}}$, the increase of ρ_{T} corresponds to a lowering of the number of sources N_{S}. The plots on the right show the fraction of the normalized average degree $\overline{k}^*/\overline{k}$.

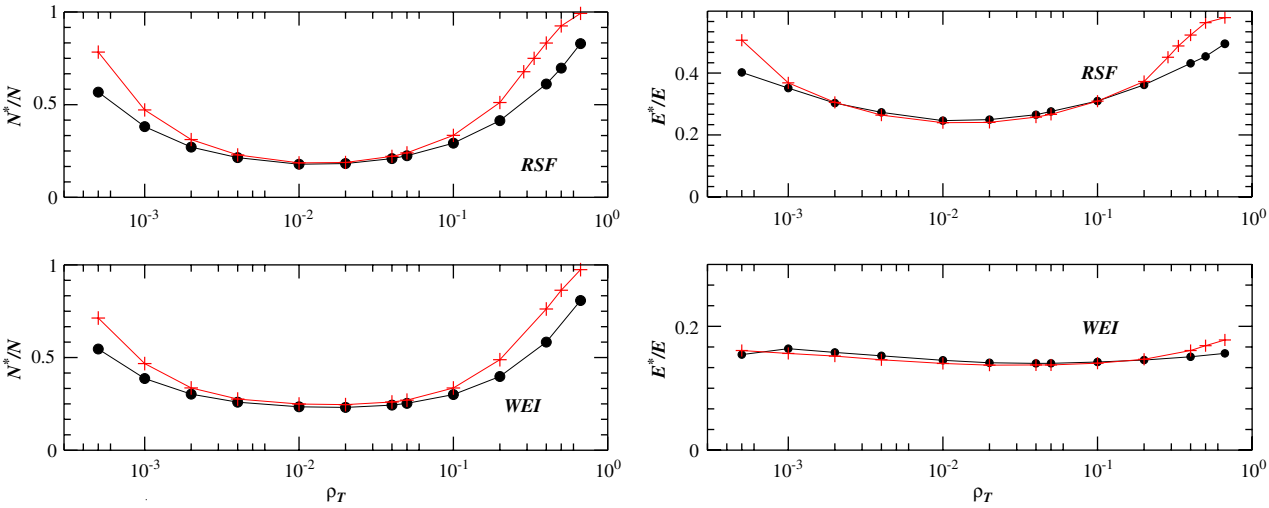

Fig. 11. Behavior as a function of ρ_T of the fraction of discovered edges and vertices in explorations with fixed ε (here $\varepsilon = 2$). The circles correspond to a random deployment of sources and targets while the crosses are obtained when sources and targets are vertices with lowest betweenness vertices.

the symmetry point. Moreover, USP 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 deployment of sources and targets. The picture, however, is more complicate if we look at other quantities in the sampled graph. In Fig. 10 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. 10 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 Section 4: 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. 11 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.

9. 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. 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 vertices, 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. Quite surprisingly, the sampling of homogeneous graphs appears more cumbersome than those of heavy-tailed graphs. Dramatic effects such as the existence of apparent power-laws, however, are found only in very peculiar cases. In general, exploration strategies provide sampled distributions with enough signatures to distinguish at the statistical level between graphs with different topologies.

This evidence might be relevant in the discussion of real data from Internet mapping projects. Indeed, data available so far indicate the presence of heavy-tailed degree distribution both at the router and AS level. In the light of the present discussion, it is very unlikely that this feature is just an artifact of the mapping strategies. The upper degree cut-off at the router and AS level runs up to 10^2 and 10^3, respectively. A homogeneous graph should have an average degree comparable to the measured cut-off and this is hardly conceivable in a realistic perspective (for instance, it would require that nine routers over 10 would have more than 100 links to other routers). In addition, the major part of mapping projects are multi-source, a feature that we have shown to readily wash out the presence of spurious power-law behavior. On the contrary, heterogeneous networks with heavy-tailed degree distributions are sampled with particular accuracy for the large degree part, generally at all probing levels. This makes very plausible, and a natural consequence, that the heavy-tail behavior observed in real mapping experiments is a genuine feature of the Internet. Furthermore, heterogeneous graphs show a striking tendency to improve the mapping efficiency at large degree vertices, while exponential graphs seem to respond in a homogeneous way independent of the degree value.

On the other hand, it is important to stress that while at the qualitative level the sampled graphs allow a discrimination of the statistical properties, at the quantitative level they might exhibit considerable deviations from the true values such as size, average degree, and the precise analytic form of the heavy-tailed degree distribution. For instance, the exponent of the power-law behavior appears to suffer from noticeable biases. In this respect, 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 [19]. 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.

Acknowledgments

We are grateful to M. Crovella, P. De Los Rios, T. Erlebach, T. Friedman, M. Latapy and T. Petermann for very useful discussions and comments. This work has been partially supported by the European Commission Fet-Open project COSIN IST-2001-33555 and contract 001907 (DELIS).

References

[1] D. Achlioptas, A. Clauset, D. Kempe, C. Moore, On the bias of traceroute sampling, or: Why almost every network looks like it has a power law, in: Proc. of the 2005 Symposium on the Theory of Computation (STOC), 21–24 May 2005. Baltimore, MD, pp. 694–703.
[2] P. Baldi, P. Frasconi, P. Smyth, Modeling the Internet and the Web: Probabilistic Methods and Algorithms, Wiley, Chichester, 2003.
[3] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[4] M. Barthélemy, Betweenness centrality in large complex networks, Eur. Phys. J. B 38 (2004) 163–168.
[5] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Soc. 25 (2) (2001) 163–177.
[6] A. Broido, K.C. Claffy, Internet topology: connectivity of IP graphs, San Diego Proc. SPIE Internat. Symp. on Convergence of IT and Communication, Denver, CO, 2001.
[7] H. Burch, B. Cheswick, Mapping the Internet, IEEE Comput. 32 (4) (1999) 97–98.
[8] G. Caldarelli, R. Marchetti, L. Pietronero, The fractal properties of internet, Europhys. Lett. 52 (2000) 386.
[9] Q. Chen, H. Chang, R. Govindan, S. Jamin, S.J. Shenker, W. Willinger, The origin of power laws in internet topologies revisited, Proc. IEEE Infocom 2002, New York, USA, 2002.
[10] A. Clauset, C. Moore, Accuracy and scaling phenomena in internet mapping, Phys. Rev. Lett. 94 (2005) 018701.
[11] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW, Oxford University Press, Oxford, 2003.
[13] P. Erdös, P. Rényi, On random graphs I, Publ. Math. Debrecen 6 (1959) 290.

[14] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, ACM SIGCOMM '99, Computers and Communications Review, Vol. 29, 1999, pp. 251–262.
[15] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.
[16] K.-I. Goh, B. Kahng, D. Kim, Universal behavior of load distribution in scale-free networks, Phys. Rev. Lett. 87 (2001) 278701.
[17] R. Govindan, H. Tangmunarunkit, Heuristics for internet map discovery, Proc. IEEE Infocom 2000, Vol. 3, IEEE Computer Society Press, Silverspring, MD, 2000, pp. 1371–1380.
[18] J.-L. Guillaume, M. Latapy, Relevance of massively distributed explorations of the internet topology: simulation results, in: Proc. of the 24th IEEE International Conference Infocom'05, 2005, Miami, USA.
[19] http://www.traceroute at home.net/
[20] Internet mapping project at Lucent Bell Labs (http://www.cs.bell-labs.com/who/ches/map/).
[21] C. Jin, Q. Chen, S. Jamin, INET: internet topology generators, Tech. Rep. CSE-TR-433-00, EECS Department, University of Michigan, 2000.
[22] A. Lakhina, J.W. Byers, M. Crovella, P. Xie, Sampling biases in IP topology measurements, Tech. Rep. BUCS-TR-2002-021, Department of Computer Sciences, Boston University, 2002.
[23] A. Medina, I. Matta, BRITE: a flexible generator of Internet topologies, Tech. Rep. BU-CS-TR-2000-005, Boston University, 2000.
[24] M. Molloy, B. Reed, A critical point for random graphs with a given degree sequence, Random Struct. Algorithms 6 (1995) 161;
M. Molloy, B. Reed, The size of the giant component of a random graph with a given degree distribution, Combinatorics, Probab. Comput. 7 (1998) 295
[25] R. Pastor-Satorras, A. Vázquez, A. Vespignani, Dynamical and correlation properties of the Internet, Phys. Rev. Lett. 87 (2001) 258701;
A. Vázquez, R. Pastor-Satorras, A. Vespignani, Large-scale topological and dynamical properties of the Internet, Phys. Rev. E 65 (2002) 066130
[26] R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet: A Statistical Physics Approach, Cambridge University Press, Cambridge, 2004.
[27] T. Petermann, P. De Los Rios, Exploration of scale-free networks—do we measure the real exponents?, Eur. Phys. J. B 38 (2004) 201–204.
[28] SCAN project at the Information Sciences Institute (http://www.isi.edu/div7/scan/).
[29] The Cooperative Association for Internet Data Analysis (CAIDA), located at the San Diego Supercomputer Center (see http://www.caida.org/home/).
[30] The National Laboratory for Applied Network Research (NLANR), sponsored by the National Science Foundation (see http://moat.nlanr.net/).
[31] Topology project, Electric Engineering and Computer Science Department, University of Michigan (http://topology.eecs.umich.edu/).
[33] W. Willinger, R. Govindan, S. Jamin, V. Paxson, S. Shenker, Scaling phenomena in the Internet: critically examining criticality, Proc. Natl. Acad. Sci. USA 99 (2002) 2573–2580.