Sampling of networks with traceroute-like probes

Alain Barrat [a,*], Ignacio Alvarez-Hamelin [a], Luca Dall'Asta [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

A large part of the recent development of the interest in complex networks has been triggered by the observation of particular characteristics of real world networks, such as the small-world properties or the heavy-tailed distributions of degrees. Many datasets are however the result of an incomplete sampling of the underlying real networks, and it has been argued that sampling procedures might introduce uncontrolled biases in the statistical properties of the sampled graph. In this paper, we explore this issue in the case of the Internet, which is generally mapped from a limited set of sources by using `traceroute`-like probes. The origin of the biases introduced by such a sampling process is investigated and related with the global topological properties of the underlying network. We complement the analytical discussion with a throughout numerical investigation of simulated mapping strategies in network models with different topologies.

Key words: Network Sampling, Traceroute, Internet exploration, Topology inference

* Corresponding author: A. Barrat, LPT, Bâtiment 210, Université de Paris-Sud, 91405 ORSAY Cedex France; email: Alain.Barrat@th.u-psud.fr; Tel: +33 1 69 15 82 22; Fax: +33 1 69 15 82 87

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 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.

This strategy has led to the obtention of various maps of the Internet [1–5] which have been used for the statistical characterization of the network. 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 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 $2 \leq \gamma \leq 2.5$ [6]. Evidence for the heavy-tailed behavior of the degree distribution has been collected in several other studies at the router and AS level [7–11] and have generated a large activity in the field of network modeling and characterization [12–16].

The obtained maps are however undoubtedly incomplete. Along with technical problems such as the instability of paths between routers and interface resolutions [17], typical mapping projects are run from relatively small sets of sources whose combined views are missing a considerable number of edges and vertices [11,18]. 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 [19–22]. Very interestingly, 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 Erdös-Rényi graph model [19,20,23]. 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.

This issue may be tackled through a mean-field statistical analysis and extensive numerical study of shortest path routed sampling, considered as the first approximation to `traceroute`-sampling (see Section 2), in different networks models. We recall in Section 3 the theoretical arguments leading to an approximate expression for the probability of edges and vertices to be detected. The analytical study provides a general understanding of which kind of topologies yields the most accurate sampling. In particular, the map accuracy depends on the underlying network *betweenness centrality* distribution; the heavier the tail the higher the statistical accuracy of the sampled graph.

Numerical investigation of maps obtained varying the number of source-target pairs on networks models with different topological properties provides support to the analytical analysis. In particular, we consider networks with degree distribution with poissonian, Weibull and power-law behavior. We study the fractions of discovered vertices and edges as a function of the degree (Section 4), stressing the agreement with the theoretical predictions, as well as the degree distributions obtained in the sampled graph (Section 5). Single source mapping processes are shown to face serious limitations: even 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.

In Section 6, 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. A region of low efficiency (less vertices and edges discovered) is found, 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 Network models and `traceroute`-like processes

In a typical `traceroute` study, active sources deployed in the network send `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 [17]. By merging the information collected on each path it is then possible to reconstruct a partial map of the network. 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, which may differ even considerably from the shortest path. Despite these local, often unpredictable path distortions, 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. For the sake of simplicity, we can thus define three selection mechanisms defining different ideal-paths that may account for some of the features encountered in real 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).
- 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. However, it turns out 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 (ver tices) $V = \{1, 2, \cdots, N\}$ and edges (links) E. Then let us define the sets of vertices $\mathcal{S} = \{i_1, i_2, \cdots, i_{N_S}\}$ and $\mathcal{T} = \{j_1, j_2, \cdots, j_{N_T}\}$ 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 densities $\rho_T = N_T/N$ 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, it is appropriate to work with the density of targets ρ_T while still considering N_S instead of the corresponding density. In many cases, an appropriate quantity representing the level of sampling of the networks is $\epsilon = N_S N_T/N$: 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 of Section 3 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 such as those created using Internet topology generators [13,12]. 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 Erdös-Rényi (ER) model [24]: 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.

Another important characteristic discriminating the topology of graphs is the clustering coefficient c_i that, giving the fraction of connected neighbors of a given node i, measures the local cohesiveness of nodes. The average clustering coefficient $C = \frac{1}{N}\sum_i c_i$ provides an indication of the global level of cohesiveness of the graph. This number is generally very small in random graphs that lack of correlations. In many real graphs however, the clustering coefficient appears to be very high and opportune models have been formulated to represent this property, both for homogeneous and heterogeneous graphs. In particular, we consider the construction algorithm proposed by Watts and Strogatz for small-world networks [28]: starting from a regular network (e.g. a one-dimensional lattice with connections to the $\overline{k}$ nearest neighbors along the chain), each link is rewired with a certain probability p. The resulting degree distribution has a shape similar to the case of Erdös-Rényi graphs, peaked around its average value. The clustering coefficient, however, is large if $p \ll 1$, making this network a typical example of clustered homogeneous network.

B) Heterogeneous graphs for which $P(k)$ is a broad distribution with heavy-tail and large fluctuations, spanning various orders of magnitude. The prototype of a scale-free graph is the growing network model by Albert and Barabási (BA) [29]. The preferential attachment mechanism (each new node is connected to m already existing nodes chosen with a probability proportional to their degree) yields a connected graph of $|V| = N$ nodes with $|E| = mN$ edges, having a power-law degree distribution $P(k) \sim k^{-\gamma}$ with $\gamma = 3$, and small clustering coefficient. Another growing model has been introduced by Dorogovtsev, Mendes and Samukhin (DMS) [30]: at each time step, a new node is introduced and connected to *the two extremities of a randomly chosen edge*, thus forming a triangle. A given node is thus in fact chosen with a probability proportional to its degree, which corresponds to the preferential attachment

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

	ER	ER	WS	BA	DMS	RSF	Weibull
N	10^4	10^4	10^4	10^4	10^4	10^4	10^4
$\|E\|$	10^5	5.10^5	10^5	4.10^4	2.10^4	22000	55000
$\overline{k}$	20	100	20	8	4	4.4	11
C	0.002	0.01	0.52	0.006	0.74	0.067	0.12
k_{max}	40	140	26	334	346	3500	2000

rule. The resulting graphs have a large clustering coefficient (≈ 0.74) along with a power-law degree distribution $P(k) \sim k^{-\gamma}$ with $\gamma = 3$.

Such graphs can be considered as particular since they are constructed with the preferential attachment mechanism, and we also consider random graphs with given broad degree distributions. 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 [8]. We have generated the corresponding random graphs by using the algorithm proposed by Molloy and Reed [31]: 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.

3 Mean-field theory of simulated mapping process

We begin our study by recalling briefly the mean-field statistical analysis of the simulated `traceroute` mapping done in [32]. The 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.

Starting from the above exact formula, it is interesting to look at the process on a statistical ground by studying the average over all possible realizations of the set $\Omega = \{\mathcal{S}, \mathcal{T}\}$, identified by $\langle \cdots \rangle$. An uncorrelation assumption allows to obtain the average discovery probability of an edge as

$$\langle \pi_{i,j} \rangle \simeq 1 - \prod_{l \neq m} \left(1 - \rho_T \rho_S \left\langle \sigma_{i,j}^{(l,m)} \right\rangle \right), \tag{2}$$

where we take advantage of neglecting correlations by replacing the average of the product of variables with the product of the averages. In the case of the ASP probing, $\left\langle \sigma_{i,j}^{(l,m)} \right\rangle$ 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 $\left\langle \sigma_{i,j}^{(l,m)} \right\rangle = x_{i,j}^{(l,m)} / \sigma^{(l,m)}$.

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

$$\prod_{l \neq m} \left(1 - \rho_T \rho_S \left\langle \sigma_{i,j}^{(l,m)} \right\rangle \right) \simeq \prod_{l \neq m} \exp\left(-\rho_T \rho_S \left\langle \sigma_{i,j}^{(l,m)} \right\rangle \right), \tag{3}$$

that inserted in Eq.(2) yields

$$\langle \pi_{i,j} \rangle \simeq 1 - \exp\left(-\rho_T \rho_S b_{ij} \right), \tag{4}$$

where $b_{ij} = \sum_{l \neq m} \left\langle \sigma_{i,j}^{(l,m)} \right\rangle$. In the case of the USP and RSP probing, the quantity b_{ij} is by definition the edge *betweenness centrality* $\sum_{l \neq m} x_{i,j}^{(l,m)} / \sigma^{(l,m)}$ [25,26], sometimes also refereed to as "load" [27] (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 $\langle \pi_{i,j} \rangle \simeq 2\rho_T \rho_S$, 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 ($N_S \sim \mathcal{O}(1)$) and their density tends to zero as N^{-1}. In this case it is more convenient to express the edge discovery probability as

$$\langle \pi_{i,j} \rangle \simeq 1 - \exp\left(-\epsilon \widetilde{b_{ij}}\right), \tag{5}$$

where $\epsilon = \rho_T N_S$ 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 of vertices, leading to the average

$$\langle \pi_i \rangle \simeq 1 - (1 - \rho_S - \rho_T) \exp\left(-\rho_T \rho_S b_i\right), \tag{6}$$

where b_i is the vertex betweenness centrality, that is limited in the interval $[0, N(N-1)]$ [25–27]. The betweenness value $b_i = 0$ holds for the leafs of the graph, i.e. vertices with a single edge, for which we recover $\langle \pi_i \rangle \simeq \rho_S + \rho_T$. 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 $\rho_S \sim \mathcal{O}(N^{-1})$ and in the large N limit we can conveniently write

$$\langle \pi_i \rangle \simeq 1 - (1 - \rho_T) \exp\left(-\epsilon \widetilde{b_i}\right), \tag{7}$$

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 vertex i discovered by our sampling process. Each edge departing from the vertex contributes proportionally to its discovery probability, yielding

$$\langle k_i^* \rangle = \sum_j \left(1 - \exp\left(-\epsilon \widetilde{b_{ij}}\right)\right) \simeq \epsilon \sum_j \widetilde{b_{ij}}. \tag{8}$$

The final expression is obtained for edges with $\epsilon \widetilde{b_{ij}} \ll 1$. 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

$$\langle k_i^* \rangle \simeq 2\epsilon + 2\epsilon \widetilde{b_i}. \tag{9}$$

Finally, the analysis allows to compute the edge redundancy $r_e(i,j)$ of an edge (i,j), defined as as the number of probes passing through the edge (i,j). This quantity is indeed written for a given set of probes and targets as

$$r_e(i,j) = \sum_{l \neq m} \left(\sum_{s=1}^{N_S} \delta_{l,i_s} \sum_{t=1}^{N_T} \delta_{m,i_t} \sigma_{i,j}^{(l,m)} \right). \tag{10}$$

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

$$\langle r_e(i,j) \rangle \simeq \sum_{l \neq m} \rho_T \rho_S \left\langle \sigma_{i,j}^{(l,m)} \right\rangle = \rho_T \rho_S b_{ij} \ . \tag{11}$$

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:

$$\langle r_n(i) \rangle \simeq 2\epsilon + \rho_S \rho_T b_i \ . \tag{12}$$

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.

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 appears through the key role played by edge and vertex betweenness in the expressions characterizing the graph discovery. The betweenness is a nonlocal 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.

4 Numerics

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 2. Let us first consider the case of homogeneous graphs (ER and WS model): 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 Eq. (5) and (7) 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 $\epsilon \gg \max\left[\overline{b}^{-1}, \overline{b_e}^{-1}\right]$. 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. 1 shows for the WS model 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

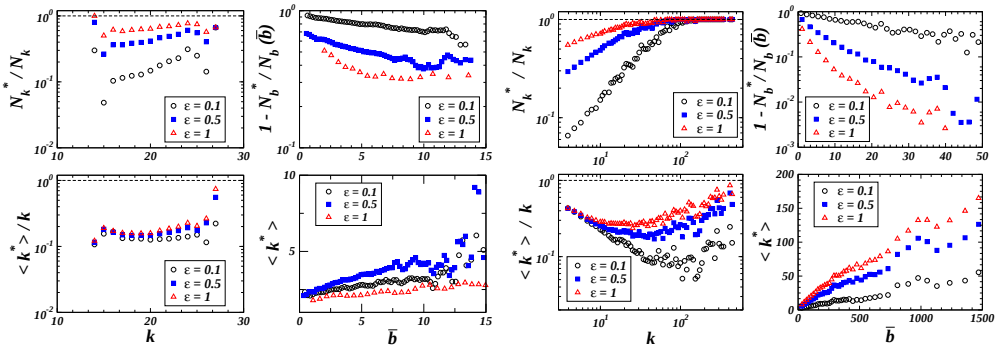

Fig. 1. Frequency N_k^*/N_k of detecting a vertex of degree k, frequency N_b^*/N_b of detecting a vertex of betweenness b and proportion of discovered edges $\langle k^* \rangle / k$ as a function of the degree and as a function of the betweenness for the WS (4 graphs on the left) and the BA (4 graphs on the right) models. The exploration setup considers $N_S = 2$ and increasing probing level ϵ obtained by progressively higher density of targets ρ_T.

a larger betweenness. 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. (9). 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 (BA, DMS, RSF and WEI models), with an appreciable fraction of vertices and edges with very high betweenness [33]. 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 [33]. Since in heavy-tailed degree distributions the allowed degree is varying over several orders of magnitude, the same occurs for the betweenness values, and the tail of the distribution is broader the broader the connectivity distribution. In such a situation, even in the case of small ϵ, vertices whose betweenness is large enough ($b_i \epsilon \gg 1$) have $\langle \pi_i \rangle \simeq 1$. Therefore all vertices with degree $k \gg \epsilon^{-1/\beta}$ will be detected with probability one. This is clearly exemplified for the BA model in Fig. 1 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 concerning $\langle k^* \rangle / k$. After an initial decay (Fig. 1) the effective discovered degree is increasing with the degree of the vertices. This qualitative feature is captured by Eq. (9) that gives $\langle k^* \rangle / k \simeq \epsilon k^{-1}(1 + \overline{b(k)})$. 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. Fig. 1 also displays the frequency N_b^*/N_b and the discovered degree of vertices with betweenness b, showing in a more direct way the qualitative agreement with the analytical predictions.

In Fig. 2 we also report the behavior of the average vertex redundancy as

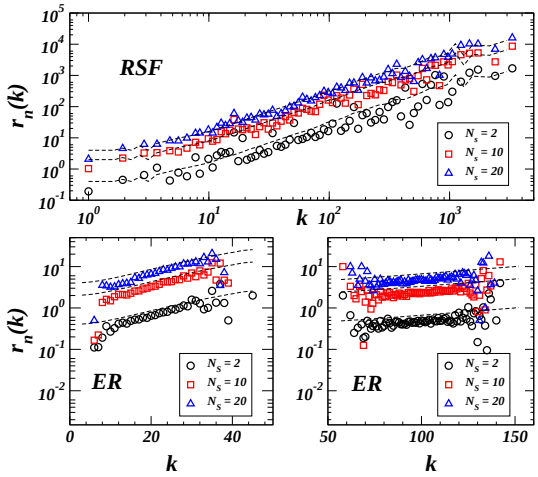

Fig. 2. Average vertex redundancy as a function of the degree k for RSF (top) and ER (bottom) model ($N = 10^4$). For the ER model, two blocks of data are plotted, for $\overline{k} = 20$ (left) and for $\overline{k} = 100$ (right) The target density is fixed ($\rho_T = 0.1$), and $N_S = 2$ (circles),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

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.

5 Degree distribution measurements

A very important quantity in the study of the statistical accuracy of the sampled graph is the degree distribution. Fig. 3 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 $N_S \geq 2$ 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 $\overline{k}$ of the underlying graph. In order to stretch the distribution range, homogeneous graphs with very large average degree

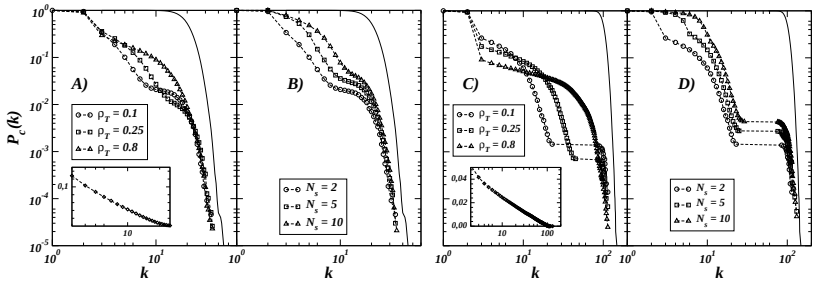

Fig. 3. Cumulative degree distribution of the sampled ER graph for USP probes. Figures A) and B) correspond to $\overline{k} = 20$, and C) and D) to $\overline{k} = 100$. Figures A) and C) show sampled distributions obtained with $N_S = 2$ and varying density target ρ_T. In the insets we report the peculiar 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. Figures 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.

$\overline{k}$ 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. Note that, in the case of RSP and ASP model, the obtained distributions are closer to the real one since they allow a larger number of discoveries.

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 [20,23]. Also in this case, the distribution cut-off is consistently determined by the average degree $\overline{k}$. 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 $n \geq 2$.

Since, in heterogeneous graphs, vertices with high degree are efficiently sampled with an effective measured degree that is rather close to the real one, 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. 4). 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 under-sampled. This undersampling can either yield an apparent change in the exponent of the degree

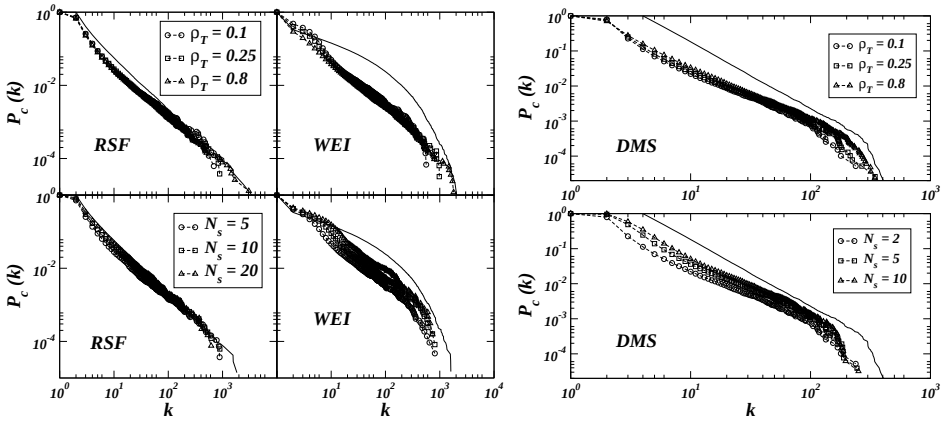

Fig. 4. Cumulative degree distributions of the sampled RSF, WEI and DMS graphs for USP probes. The top figures show sampled distributions obtained with $N_S = 5$ and varying density target ρ_T. The figures on the bottom correspond to $\rho_T = 0.25$ and varying number of sources N_S. The solid lines are the degree distributions of the underlying graph.

distribution (as also noticed in [21] for single source experiments), or, if N_S is small, yield a power-law like distribution for an underlying Weibull distribution. Furthermore, as Fig. 4 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 Sec. 7.

6 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

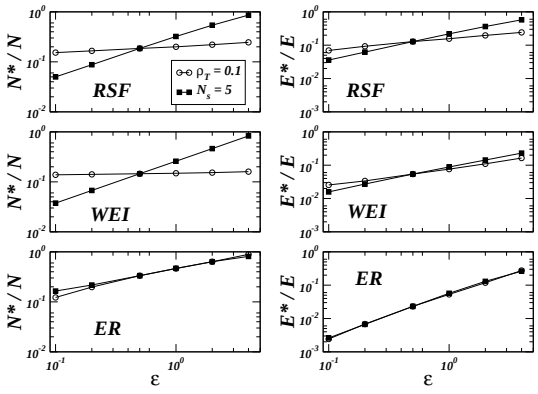

Fig. 5. Behavior of the fraction of discovered nodes and edges in explorations with increasing ϵ, for RSF, WEI and ER graphs. For each underlying graph studied we report two curves corresponding to larger ϵ achieved by increasing the target density ρ_T at constant $N_S = 5$ (squares) or the number of sources N_S at constant $\rho_T = 0.1$ (circles). Curves similar to ER are obtained for WS, and to RSF for BA and DMS.

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. 5 we report the proportion of discovered nodes and 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 density 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.

In Fig. 6, we report the behavior of E^*/E and N^*/N at fixed ϵ and varying N_S and ρ_T. Very interestingly, the curves show a structure allowing for local minima and maxima in the discovered portion of the underlying graph: 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

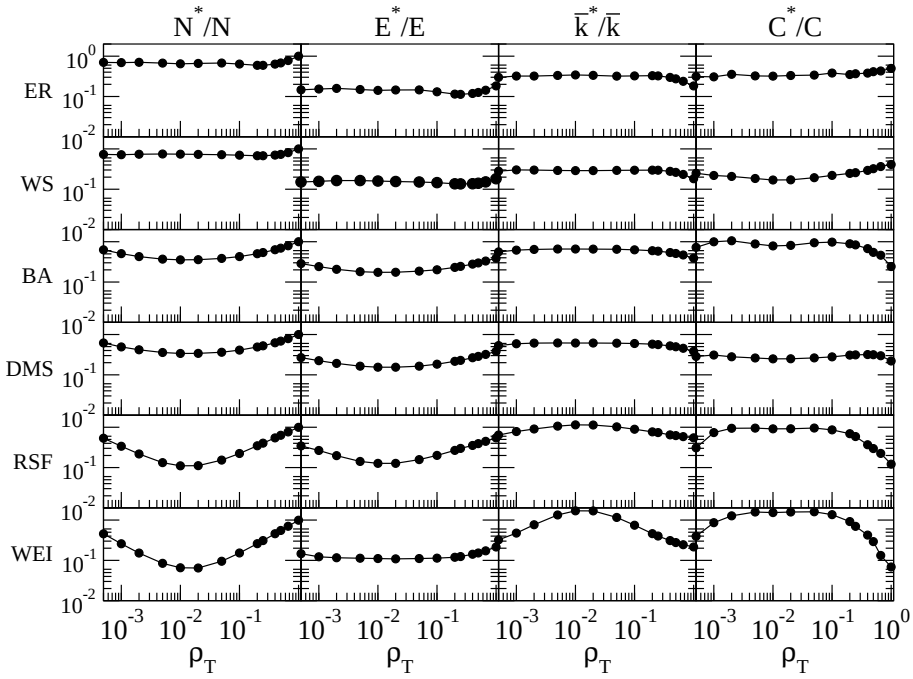

Fig. 6. Explorations with fixed ϵ (here $\epsilon = 2$): Behavior as a function of ρ_T of the fraction of discovered vertices N^*/N and edges E^*/E, of the normalized average degree $\overline{k}^*/\overline{k}$ and of the fraction of the normalized average clustering coefficient C^*/C. Since $\epsilon = \rho_T N_S$, the increase of ρ_T corresponds to a lowering of the number of sources N_S.

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.6 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.6 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. This implies that at the symmetry point the exploration discovers less edges than in other setups, however, achieving a more efficient sampling of the effective degree for the discovered vertices. A similar problem is obtained by studying the behavior of the ratio C^*/C between the clustering coefficient of the sampled and the underlying graphs: 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 3: 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. The usual correlation between connectivity and betweenness thus indicates that the exploration of a real network could be improved by a massive deployment of sources using low-connectivity vertices.

7 Conclusions and outlook

The rationalization of the sampling 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 between homogeneous and heavy-tailed topologies. 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 indicate the presence of heavy-tailed degree distribution both at the router and AS level. The present discussion indicates that 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, which is hardly conceivable in a realistic perspective (for instance, it would require that nine routers over ten would have more than 100 links to other routers). In addition, the major part of mapping projects are multi-source, a feature that readily washes 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.

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 [34,35]. 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] 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/).

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

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

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

[6] Faloutsos M, Faloutsos P, Faloutsos C: On Power-law Relationships of the Internet Topology. ACM SIGCOMM '99, Comput Commun Rev 1999; 29: 251-262.

[7] Govindan R, Tangmunarunkit H: Heuristics for Internet Map Discovery. Proc. of IEEE Infocom 2000; Volume 3, IEEE Computer Society Press: 1371–1380.

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

[9] Caldarelli G, Marchetti R, Pietronero L: The Fractal Properties of Internet. Europhys Lett 2000; 52: 386.

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

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

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

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

[14] Dorogovtsev SN, Mendes JFF: Evolution of networks: From biological nets to the Internet and WWW, Oxford University Press, Oxford, 2003.

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

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

[17] Burch H, Cheswick B: Mapping the internet. IEEE computer 1999; 32: 97-98.

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

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

[20] Clauset A, Moore C: Accuracy and Scaling Phenomena in Internet Mapping. Phys Rev Lett 2005; 94: 018701.

[21] Petermann T, De Los Rios P: Exploration of Scale-Free Networks - Do we measure the real exponents? Eur Phys J B 2004; 38: 201-204.

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

[23] Achlioptas D, Clauset A, Kempe D, Moore C: On the Bias of Traceroute Sampling; or, Power-law Degree Distributions in Regular Graphs. cond-mat/0503087, to appear in STOC 2005.

[24] Erdös P, Rényi P: On random graphs I. Publ Math Debrecen 1959; 6: 290.

[25] Freeman LC: A Set of Measures of Centrality Based on Betweenness. Sociometry 1977; 40: 35-41.

[26] Brandes U: A Faster Algorithm for Betweenness Centrality. J Math Soc 2001; 25: 163-177.

[27] Goh K-I, Kahng B, Kim D: Universal Behavior of Load Distribution in Scale-Free Networks. Phys Rev Lett 2001; 87: 278701.

[28] Watts DJ, Strogatz SH: Collective dynamics of small-world networks. Nature 1998; 393: 440-442.

[29] Barabási A-L, Albert R: Emergence of scaling in random networks. Science 1999; 286: 509-512.

[30] Dorogovtsev SN, Mendes JFF, Samukhin AN: Size-dependent degree distribution of a scale-free growing network. Phys Rev E 2001; 63: 062101.

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

[32] Dall'Asta L, Alvarez-Hamelin I, Barrat A, Vázquez A, Vespignani A: Statistical theory of Internet exploration. Phys Rev E 2005; 71: 036135.

[33] Barthélemy M: Betweenness Centrality in Large Complex Networks. Eur Phys J B 2004; 38: 163-168.

[34] http://www.netdimes.org/ ; Shavitt Y, Shir E: DIMES: Let the Internet Measure Itself. preprint cs.NI/0506099.

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