Statistical theory of Internet exploration

Luca Dall'Asta,[1] Ignacio Alvarez-Hamelin,[1,2] Alain Barrat,[1] Alexei Vázquez,[3] and Alessandro Vespignani[1,4]
[1]*Laboratoire de Physique Théorique, Bâtiment 210, Université de Paris-Sud, 91405 ORSAY Cedex, France*
[2]*Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, C 1063 ACV Buenos Aires, Argentina*
[3]*Nieuwland Science Hall, University of Notre Dame, Notre Dame, Indiana 46556, USA*
[4]*School of Informatics and Department of Physics, Indiana University, Bloomington, Indiana 47408, USA*

(Received 19 October 2004; published 24 March 2005)

The general methodology used to construct Internet maps consists in merging all the discovered paths obtained by sending data packets from a set of active computers to a set of destination hosts, obtaining a graphlike representation of the network. This technique, sometimes referred to as Internet tomography, spurs the issue concerning the statistical reliability of such empirical maps. We tackle this problem by modeling the network sampling process on synthetic graphs and by using a mean-field approximation to obtain expressions for the probability of edge and vertex detection in the sampled graph. This allows a general understanding of the origin of possible sampling biases. In particular, we find a direct dependence of the map statistical accuracy upon the topological properties (in particular, the *betweenness centrality* property) of the underlying network. In this framework, it appears that statistically heterogeneous network topologies are captured better than the homogeneous ones during the mapping process. Finally, the analytical discussion is complemented with a thorough numerical investigation of simulated mapping strategies in network models with varying topological properties.

DOI: 10.1103/PhysRevE.71.036135 PACS number(s): 89.75.−k, 89.20.Hh, 89.20.Ff

I. INTRODUCTION

In recent years a considerable research effort has been focused on the field of complex networks [1–3]. The main reason for this effort finds its rationale in the very pervasive presence of biological, social, or technological structures that can be described using the paradigm of complex networks. At a very abstract level, a network is a system composed of many elementary agents (nodes) cooperating via relations or interactions between them (links). The physical Internet is one of the most common examples of complex networks in the real society. Its growing structure is the result of competitive and cooperative processes, in which individual choice, optimization criteria, and policy-driven strategies cooperate with the lack of any centralized control in determining the self-organized evolution of the system [4,5]. All these factors lead to the formation of a complex structure, whose fabric and topology is largely unknown. In the absence of accurate Internet maps many research groups have started large scale projects aimed at the collection of data on the topology and structure of this network of networks [6–10]. Investigations can be made 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 or ASs and their physical connections are, respectively, the vertices and edges of a graph. 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 routing tables (the so-called border gateway protocol [BGP] tables) [5–10].

The importance of traceroutelike mapping processes resides in their simplicity and generality. The traceroute command sends probes (data packets) toward a certain Internet node (Internet provider address), providing the addresses of the traversed nodes. The merging of the discovered paths allows the construction of a graphlike representation of the Internet whose vertices are routers or ASs. By using tracerouteike mapping processes, a number of research groups have generated maps of the Internet [6–10] that have been used for the statistical characterization of the network properties. The obtained maps show that the undirected *sparse* graph representing the Internet is a *small world*, an essential property for the efficient functioning of an information network. More strikingly, many studies have reported evidence for a heavy-tailed behavior of the Internet degree distribution: in particular, in [11], a power-law degree distribution $P(k) \sim k^{-\gamma}$ with $2 \leqslant \gamma \leqslant 2.5$ has been found. Several other studies have collected data from Internet explorations, all confirming a broad behavior of the degree distribution, at both the router and AS level [12–16]. The evidence for a very heterogeneous topology of the Internet, prompting the inadequacy of the standard paradigm of homogeneous networks, has thus generated a large activity in the field of network modeling [2,4,5,17,18].

Despite the flexibility of traceroute-driven strategies, the obtained maps are undoubtedly incomplete. In addition to factors causing path distortion and other subtle technical problems [19], the relatively small number of sources from which the mapping projects are usually run allows combined views missing a considerable fraction of edges and vertices [16,20]. In particular, the various spanning trees are especially missing those links that belong to transversal paths with respect to the shortest paths toward the targets (the so called *lateral connectivity*). Moreover, they sample more frequently nodes and links that 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 [21–24]. In the case of a single source, it has been shown that apparent degree distributions with heavy tails may be observed for the sampled graph even if the underlying real graph is homogeneous (such as in the classic Erdös-Rényi graph model) [21,22], or that the measure of the exponents of the degree distribution can be biased [22,23]. 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.

While traceroute probes take into account traffic loads, differences of bandwidth, policy strategies, failures, and other factors that can affect the actual path chosen by packets, the simplest model of traceroute exploration amounts to consider the collection of shortest paths for a source-target pair. Indeed, shortest path routing can be considered as a first approximation to the real probing path and the merging of several of these views an approximation to the mapping process. In this work we focus on the tight relation between statistical (topological) properties of a network and tracer-outelike mapping strategies based upon shortest path routing, with the purpose of understanding how the different topologies respond to the sampling process and what are their characteristic signatures in terms of statistical quantities. We tackle the problem by performing a mean-field statistical analysis and extensive numerical experiments of shortest path routed traceroutelike sampling in different network models. We find an approximate expression for the probability of edges and vertices to be detected that exploits the dependence upon the number of sources and targets and the topological properties (sparseness, betweenness) of the networks. This expression allows the understanding, at a qualitative level, of the efficiency of exploration methods by changing the number of probes imposed on the graph. 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 distribution; the heavier the tail, the higher the statistical accuracy of the sampled graph.

We substantiate our analytical finding with a thorough analysis of maps obtained by varying the number of source-target pairs on network models with different topological properties. The 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 a consistent increase in the accuracy of the obtained maps, where the statistical degree distributions are qualitatively discriminated even at low values of target density. A detailed discussion of the behavior of the degree distribution and other statistical quantities, as a function of the number of targets and sources, is provided for sampled graphs with different topologies, and compared with the insight obtained by analytical means.

The paper is structured as follows. In Sec. II we discuss the theoretical model of traceroutelike processes. In Sec. III, a mean-field statistical analysis of the model is developed, in order to obtain analytical predictions of the discovery probabilities. A throughout numerical exploration of several networks with different topological properties is provided in Sec. IV, stressing the agreement between analytical predictions and numerical results.

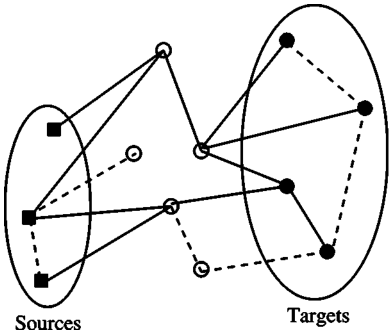

FIG. 1. Illustration of the traceroutelike procedure. Shortest paths between the set of sources and the set of destination targets are discovered (shown in full lines) while other edges are not found (dashed lines). Note that not all shortest paths are found since the "unique shortest path" procedure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

II. MODELING THE TRACEROUTE DISCOVERY OF UNKNOWN NETWORKS

As sketched in the Introduction, in a typical traceroute study, a set of active sources deployed in the network runs traceroute probes to a set of destination nodes. Each probe collects information on all the nodes and edges visited along the path connecting the source to the destination, allowing the discovery of the network [19]. By merging the information collected on each path, it is possible to reconstruct a partial map of the network (see Fig. 1). While in the Internet many factors, including commercial agreement and administrative routing policies, contribute to determine the actual path, it is clear that, to a first approximation, the route obtained by traceroutelike probes is the shortest path between the two nodes. This assumption, however, is not sufficient for a proper definition of a traceroute model in that equivalent shortest paths between two nodes may exist. In the presence of a degeneracy of shortest paths we must therefore specify the traceroute model by providing a resolution algorithm for the selection of shortest paths.

For the sake of simplicity, we can define three selection mechanisms.

(1) The unique shortest path (USP) probe. In this case the shortest path route selected between two nodes i and j is always the same independently of the source S and target T (the path being initially chosen at random among all the equivalent ones).

(2) The random shortest path (RSP) probe. The shortest path between any node pair is chosen randomly among the set of equivalent shortest paths. This might mimic the effects of traffic congestion and administrative policies that can make independent the paths among pairs of nodes.

(3) The all shortest paths (ASP) probe. This procedure 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 equivalent paths are discovered in different runs.

Actual traceroute probes contain a mixture of the three mechanisms defined above, though we do not attempt to account for all the subtleties that real studies encounter. Each traceroute model provides a test of the possible biases and we will see that the different mechanisms have only little influence on the general picture emerging from our results. On the other hand, it is intuitive to recognize that the USP model represents the worst case among the three different methods, since it yields the minimum number of discoveries. In this perspective, even real mapping should provide very likely a more optimistic scenario than those determined by the USP case. For this reason, if not otherwise specified, we will report the USP data to illustrate the general features of our synthetic exploration.

More formally, the experimental setup for our simulated traceroute mapping is the following. Let $G=(V,E)$ be a sparse undirected graph with N nodes; we 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 the path connecting each source-target pair according to the USP method. 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 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 N_S instead of the corresponding density. 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 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

$$\epsilon=\frac{N_S N_T}{N}, \tag{1}$$

that measures the density of probes imposed on 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 what 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.

III. MEAN-FIELD THEORY OF THE DISCOVERY BIAS

By means of the following mean-field statistical analysis of the simulated traceroute mapping, we provide a statistical estimate for the probability of edge and node 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 assumes the value 1 if the edge (i,j) belongs to the path selected by the traceroute model between nodes l and m, and 0 otherwise. For a given set $\Omega=\{\mathcal{S},\mathcal{T}\}$, the characteristic function that indicates if a given edge (i,j) is discovered and belongs to the sampled graph can then be written as

$$\pi_{i,j}=1-\prod_{l\neq m}\left(1-\sum_{s=1}^{N_S}\delta_{l,i_s}\sum_{t=1}^{N_T}\delta_{m,i_t}\sigma_{i,j}^{(l,m)}\right). \tag{2}$$

This function is simply $\pi_{i,j}=1$ if the edge (i,j) belongs to at least one of the paths connecting the source-target pairs, and 0 otherwise. The average over all possible realizations of the set $\Omega=\{\mathcal{S},\mathcal{T}\}$ gives us the statistical counterpart of the characteristic function, that is, the discovery probability. In the following, we will make use of an uncorrelation assumption that allows an explicit approximation for the discovery probability. Neglecting correlations between the paths generated by different source-target pairs, the discovery probability is thus obtained by considering the edge in an average effective medium of sources and targets homogeneously distributed in the network. In this approximation, the average of the product can be replaced by the product of the averages, and recalling that each node i has, on average, a probability to be a source or a target proportional to their respective densities,

$$\left\langle\sum_{t=1}^{N_T}\delta_{i,i_t}\right\rangle=\rho_T \quad\text{and}\quad \left\langle\sum_{s=1}^{N_S}\delta_{i,i_s}\right\rangle=\rho_S, \tag{3}$$

the *average discovery probability of an edge* is

$$\begin{aligned}\langle\pi_{i,j}\rangle&=1-\left\langle\prod_{l\neq m}\left(1-\sum_{s=1}^{N_S}\delta_{l,i_s}\sum_{t=1}^{N_T}\delta_{m,i_t}\sigma_{i,j}^{(l,m)}\right)\right\rangle\\&\simeq 1-\prod_{l\neq m}(1-\rho_T\rho_S\langle\sigma_{i,j}^{(l,m)}\rangle).\end{aligned} \tag{4}$$

This expression simply states that each possible source-target pair is weighted in the average with the product of the probability that the end nodes are a source and a target. The realization average of $\langle\sigma_{i,j}^{(l,m)}\rangle$ is very simple in the uncorrelated picture, depending only on the kind of probing model. In the case of the ASP method, all shortest paths are discovered, so that $\langle\sigma_{i,j}^{(l,m)}\rangle$ is just 1 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, the situation is slightly different since only one of the possibly multiple shortest paths between l and m is discovered. 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 going through (i,j), it is then clear that 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 $\langle\sigma_{i,j}^{(l,m)}\rangle\leq 1$, we have

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

which inserted in Eq. (4) yields

$$\langle \pi_{i,j}\rangle \simeq 1-\prod_{l\neq m}\left[\exp(-\rho_T\rho_S\langle\sigma_{i,j}^{(l,m)}\rangle)\right]=1-\exp(-\rho_T\rho_S b_{ij}), \tag{6}$$

where $b_{ij}=\Sigma_{l\neq m}\langle\sigma_{i,j}^{(l,m)}\rangle$. In the case of the USP and RSP, the quantity b_{ij} is by definition the edge betweenness centrality $\Sigma_{l\neq m}x_{i,j}^{(l,m)}/\sigma^{(l,m)}$ [25,26], sometimes also refereed to as the "load" [27] (in the case of ASP, it is a closely related quantity). The betweenness, a classical nonlocal measure of the *centrality* of an edge or vertex in the graph, can be seen, in this context, as a measure of the traffic load that goes through an edge or vertex, if the shortest path is used as defining the optimal path between pairs of vertices.

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 vertices with minimum betweenness $b_{ij}=2$ we have

$$\langle\pi_{i,j}\rangle \simeq 2\rho_T\rho_S, \tag{7}$$

which 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 a large majority of edges with high betweenness, the discovery probability approaches 1 and we can reasonably expect to have a fair sampling of the network.

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 $[N_S \sim O(1)]$ and their density tends to zero as N^{-1}. In this case it is more convenient to express the edge discovery probability as

$$\langle\pi_{i,j}\rangle \simeq 1-\exp(-\epsilon\overline{b}_{ij}), \tag{8}$$

where $\epsilon=\rho_T N_S$ is the density of probes imposed to the system and the rescaled betweenness $\overline{b}_{ij}=N^{-1}b_{ij}$ 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 $\langle\pi_{i,j}\rangle\sim O(N^{-1})$, for any finite value of ϵ. This readily tells us 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 $\langle\pi_i\rangle$ of a vertex i. For each source-target set Ω we have that

$$\pi_i = 1-\left(1-\sum_{s=1}^{N_S}\delta_{i,i_s}-\sum_{t=1}^{N_T}\delta_{i,i_t}\right)\times\prod_{l\neq m\neq i}\left(1-\sum_{s=1}^{N_S}\delta_{l,i_s}\sum_{t=1}^{N_T}\delta_{m,i_t}\sigma_i^{(l,m)}\right). \tag{9}$$

where $\sigma_i^{(l,m)}=1$ if the vertex i belongs to the path selected by the traceroute model between nodes l and m, and 0 otherwise. In this formula, it has been considered that vertices belonging to the set of sources or targets are discovered with probability 1. The second term on the right hand side, therefore, expresses that the vertex i does not belong to the set of sources and targets and is not discovered by any selected path between source-target pairs. By using the same mean-field approximation as previously, the average vertex discovery probability reads as

$$\langle\pi_i\rangle \simeq 1-(1-\rho_S-\rho_T)\prod_{l\neq m\neq i}(1-\rho_T\rho_S\langle\sigma_i^{(l,m)}\rangle). \tag{10}$$

As for the case of the edge discovery probability, the average considers all possible source-target pairs weighted with probability $\rho_T\rho_S$. 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 $\rho_T\rho_S\ll 1$, by using the same approximations used to obtain Eq. (6), we obtain

$$\langle\pi_i\rangle \simeq 1-(1-\rho_S-\rho_T)\exp(-\rho_T\rho_S b_i), \tag{11}$$

where $b_i=\Sigma_{l\neq m\neq i}\langle\sigma_i^{(l,m)}\rangle$. For the USP and RSP cases, $b_i = \Sigma_{l\neq m\neq i}x_i^{(l,m)}/\sigma^{(l,m)}$ is the vertex betweenness centrality which is limited in the interval $[0,N(N-1)]$ [25–27]. For instance, the leaves of the graph are dangling ends discovered only if they are either a source or a target themselves; they have betweenness value $b_i=0$ and, indeed, we recover $\langle\pi_i\rangle\simeq\rho_S+\rho_T$.

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

$$\langle\pi_i\rangle \simeq 1-(1-\rho_T)\exp(-\epsilon\overline{b}_i), \tag{12}$$

where we have neglected terms of order $O(N^{-1})$ and the rescaled betweenness $\overline{b}_i=N^{-1}b_i$ is now defined in the interval $[0,N-1]$. This expression points out that the probability of vertex discovery is favored by the use 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 node i discovered by our sampling process. Each edge departing from the vertex will contribute proportionally to its discovery probability, yielding

$$\langle k_i^*\rangle = \sum_j\left[1-\exp(-\epsilon\overline{b}_{ij})\right]\simeq\epsilon\sum_j\overline{b}_{ij}. \tag{13}$$

The final expression is obtained for edges with $\epsilon\overline{b}_{ij}\ll 1$. In this case, the sum over all neighbors of the edge betweenness is simply related to the vertex betweenness as $\Sigma_j b_{ij}=2(b_i+N-1)$, 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 end point. This finally yields

$$\langle k_i^*\rangle \simeq 2\epsilon+2\epsilon\overline{b}_i. \tag{14}$$

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 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.

IV. NUMERICAL EXPLORATION OF GRAPHS

Let us consider a sparse undirected graph, denoted by $G=(V,E)$. We will consider two main classes: (i) homogeneous and (ii) heterogeneous graphs. Graphs are considered to be *homogeneous* if the degree distribution $P(k)$ is peaked around its average value $\overline{k}$. This average is then meaningful and typical of any given vertex. On the contrary, *heterogeneous* graphs display degree values ranging over various orders of magnitude, and the average value is not representative or typical (for example, the maximum value of the degree, $k_{\max}$, is much larger than $\overline{k}$). As a prototype of heterogeneous graphs, we consider the class of *scale-free graphs*, for which $P(k)$ has a heavy tail decaying as a power law $P(k)\sim k^{-\gamma}$; such graphs are very heterogeneous, with large fluctuations of the degree, characterized by a variance of the degree distribution diverging with the size of the network.

Another important characteristic discriminating the topology of graphs is the clustering coefficient c_i which, giving the fraction of connected neighbors of a given node i, measures the local cohesiveness of nodes. The average clustering coefficient $C=(1/N)\Sigma_i c_i$ provides an indication of the global level of cohesiveness of the graph. This number is generally very small in random graphs that lack correlations. In many real graphs, however, the clustering coefficient appears to be very high and opportune models have been formulated to represent this property, for both homogeneous and heterogeneous graphs. In the following sections we will make use of those models that can be considered typical examples of the various classes. The numerical procedure is in all cases the following: (i) we consider a graph with given topological properties; (ii) we choose at random N_S vertices as sources and N_T vertices as targets; (iii) we compute the shortest paths between sources and targets; (iv) the properties of the graph obtained by the merging of the shortest paths are analyzed and compared to those of the original graph, in particular to test the predictions of the mean-field analysis.

A. Sampling homogeneous graphs

Our first set of simulations considers underlying graphs with homogeneous connectivity; namely, the Erdös-Rényi (ER) and the Watts-Strogatz (WS) models.

The classical Erdös-Rényi model [28] is a typical example of a homogeneous graph, with degree distribution following a Poisson law, and very small clustering coefficient (of order $1/N$). Since an ER graph can consist of more than one connected component, we consider only the largest of these components. Another homogeneous graph can be obtained with the construction algorithm proposed by Watts and Strogatz for small-world networks [29]: 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 the rewiring probability $p\ll 1$, making this network a typical example of a clustered homogeneous network. As for the ER case, it is possible to obtain graphs consisting of more than one connected component; in this case we use the largest of these components.

TABLE I. Main characteristics of the graphs used in the numerical exploration.

	ER	ER	WS	BA	DMS
N	10^4	10^4	10^4	10^4	10^4
$\lvert E\rvert$	10^5	5×10^5	10^5	4×10^4	2×10^4
$\overline{k}$	20	100	20	8	4
C	0.002	0.01	0.52	0.006	0.74
$k_{\max}$	40	140	26	334	346

We have used networks with $N=10^4$ nodes, $\overline{k}=20$ unless otherwise specified; for the WS model, $p=0.1$ has been taken (see Table I). Each measurement is averaged over ten realizations. For both models, and similarly to the degree distribution, the vertex and edge betweenness distributions are peaked around their average values $\overline{b}$ and $\overline{b}_e$, respectively. The node betweenness cumulative distribution is reported in Fig. 2, confirming the narrowness of the values interval around the characteristic value $\overline{b}$, with maximal values much smaller than N, and increasing only slowly with N. Since a large majority of vertices and edges will have a betweenness very close to the average value, we can use Eqs. (8) and (12) to estimate the order of magnitude of probes that allows a

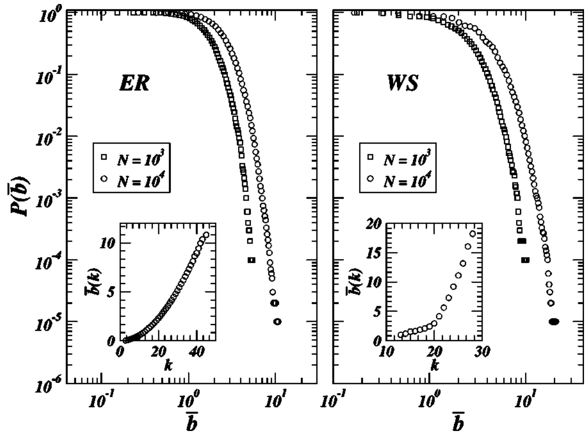

FIG. 2. Cumulative distribution of the average node betweenness $\overline{b}$ in the ER and WS graph models. The inset (in linear-linear scale) shows the behavior of the average node betweenness as a function of the degree k.

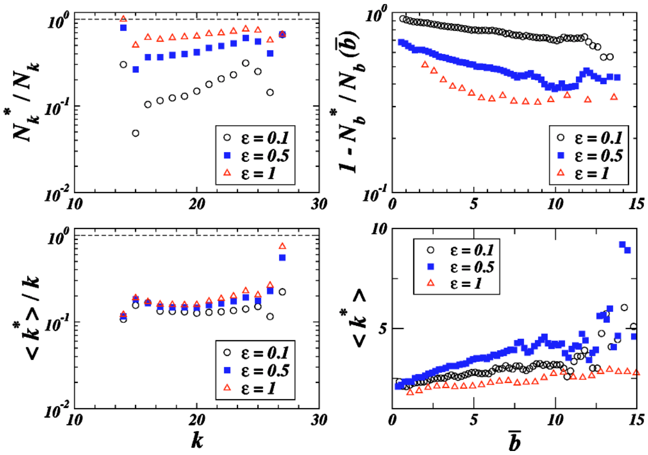

FIG. 3. Frequency N_k^*/N_k of detecting a vertex of degree k (top left) and proportion of discovered edges $\langle k^* \rangle / k$ (bottom left) as a function of the degree in WS graphs. The figures on the right show the frequency N_b^*/N_b of detecting a vertex of betweenness b and the effective average degree $\langle k^* \rangle$ as a function of the betweenness centrality, in order to provide a direct comparison with the predictions of Eqs. (12) and (14). The exploration setup considers $N_S=2$ and increasing probing level ϵ obtained by progressively higher density of targets ρ_T.

fair sampling of the graph. Indeed, both $\langle \pi_{i,j} \rangle$ and $\langle \pi_i \rangle$ tend to 1 if $\epsilon \gg \max[\overline{b}^{-1}, \overline{b}_e^{-1}]$. In this limit all edges and vertices will have probability to be discovered very close to 1.

At lower value of ϵ, obtained by varying ρ_T and N_S, the underlying graph is only partially discovered. We first study the behavior of the fraction N_k^*/N_k of discovered vertices of degree k, N_k being 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. In Fig. 3 we report the behavior of these quantities as a function of k for the WS model (a similar behavior is obtained for ER graphs). The fraction N_k^*/N_k naturally increases by augmenting the density of targets and sources, and it is slightly increasing for larger degrees. The latter behavior can be easily understood by noticing that vertices with larger degree have on average a larger betweenness $b(k)$ (see inset of Fig. 2). By using Eq. (12) we have that $N_k^*/N_k \sim 1 - \exp[-\epsilon \overline{b(k)}]$, obtaining the observed increase at large k. On the other hand, the range of variation of degree and betweenness 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. (14) stating that $\langle k^* \rangle / k \simeq \epsilon k^{-1}[1 + \overline{b(k)}]$. Indeed the initial decrease of $\langle k^* \rangle / k$ is finally compensated by the increase of $\overline{b(k)}$.

A very important quantity in the study of the statistical accuracy of the sampled graph is the degree distribution. In Fig. 4 we show the cumulative degree distribution $P_c(k^* > k)$ of the sampled graph defined by the ER model for increasing number of sources. The sampled distributions are only approximating the genuine distribution. In particular, for $N_S=1$, a power law is obtained (inset of Fig. 4), in striking contrast with the genuine degree distribution of the real graph, as analytically shown by Clauset and Moore [22]. However, strong deviations from this power law appear as soon as $N_S \geq 2$, and the obtained distributions are far from a true heavy-tail distribution at any appreciable level of probing. Indeed, the distribution runs generally over a small range of degrees, with a cutoff 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 $\overline{k}$ must be considered; however, other distinctive spurious effects appear in this case as soon as $N_S \geq 2$. 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. 4). The very same behavior is obtained in the case of the WS model. Finally, in the case of RSP and ASP traceroute models, we observe that the obtained distributions are closer to the real one since they allow a larger number of discoveries.

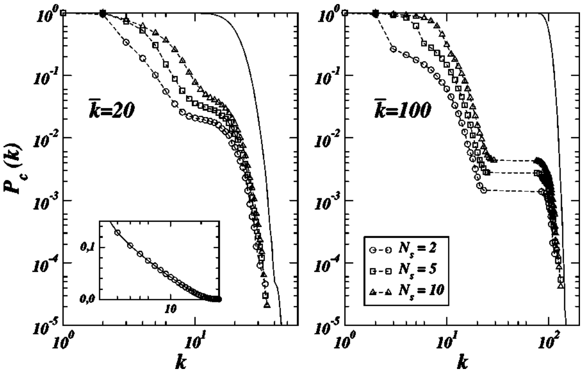

FIG. 4. Cumulative degree distribution of the sampled ER graph with $\overline{k}=20$ and 100, for USP probes. The figure shows sampled distributions obtained with $\rho_T=0.1$ and varying number of sources N_S. The solid lines are the degree distributions of the underlying graphs. In the inset we report the peculiar case $N_S=1$ which provides an apparent power-law behavior with exponent -1 at all values of ρ_T. The inset is in linear-log scale to show the logarithmic behavior of the corresponding cumulative distribution.

Only the particular case[1] of $N_S=1$ yields for the sampled distribution an apparent scale-free behavior with slope -1 (for all target densities ρ_T [22]). The distribution cutoff is then 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 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 $n \geq 2$.

B. Sampling heterogeneous graphs

In this section, we extend the analysis made for homogeneous graphs to the case of highly heterogeneous graphs. As

[1] 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.

typical examples, we consider the Barabási-Albert (BA) and the Dorogovtsev, Mendes, and Samukhin (DMS) models, which are both scale-free graph models.

The prototype of a scale-free graph is the original growing network model by Albert and Barabási [30]. The preferential attachment mechanism (each new node is connected to already existing nodes chosen with a probability proportional to their degree) yields a connected graph with a power-law degree distribution and small clustering coefficient. Another growing model has been introduced by Dorogovtsev, Mendes, and Samukhin [31]: 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 rule. The resulting graphs have a large clustering coefficient (≈ 0.74) along with a power-law degree distribution.

We have used networks of size $N=10^4$ with $\overline{k}=8$ for the BA and $\overline{k}=4$ for the DMS model, and averaged each measurement over ten realizations (see Table I). Both models have a scale-free distribution $P(k)\sim k^{-\gamma}$ with $\gamma=3$. Since the degree distribution is heavy tailed with fluctuations diverging logarithmically with the graph size, the average degree $\overline{k}$, though well defined, is not a typical value in the network and there is an appreciable probability of finding vertices with very high degree. Analogously, the betweenness distribution is heavy tailed [27,32], allowing for an appreciable fraction of vertices and edges with very high betweenness. In particular it is possible to show that 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 [32]. Since in a heavy-tailed distribution the allowed degree is varying over several orders of magnitude, the same will occur for the betweenness values. In such a situation, even in the case of small ϵ, vertices whose betweenness is large enough $[\overline{b(k)}\epsilon \gg 1]$ have $\langle \pi_i \rangle \simeq 1$. Therefore, all vertices with degree $k \gg \epsilon^{-1/\beta}$ will be detected with probability 1. This is clearly visible in Fig. 5, where the discovery probability N_k^*/N_k of vertices with degree k saturates to 1 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 (see Fig. 5) the effective discovered degree is increasing with the degree of the vertices. This qualitative feature is captured by Eq. (14) which gives $\langle k^* \rangle / k \simeq \epsilon k^{-1}[1+\overline{b(k)}]$. After an initial decay the term $k^{-1}\overline{b(k)}\sim k^{\beta-1}$ takes over and the effective discovered degree approaches the real degree k. Figure 5 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 of Eqs. (12) and (14). It is worth stressing that the results obtained for the DMS model show the very same behavior as those obtained in the case of the BA model.

It is evident from the previous discussions that, in scale-free 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 BA and DMS graphs. In Fig. 6 we report the degree distribution obtained for the DMS model. Similar plots are obtained in the case of the BA model with the same level of probing. Although both underlying DMS and BA 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 val-

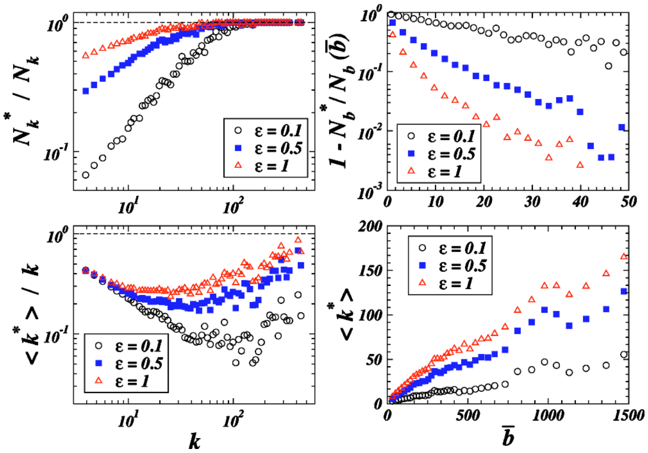

FIG. 5. Frequency N_k^*/N_k of detecting a vertex of degree k (top left) and proportion of discovered edges $\langle k^* \rangle / k$ (bottom left) as a function of the degree in the BA model. The figures on the right show the frequency N_b^*/N_b of detecting a vertex of betweenness b and the effective average degree $\langle k^* \rangle$ as a function of the betweenness centrality, in order to provide a direct comparison with the predictions of Eqs. (12) and (14). The exploration setup considers $N_S=2$ and increasing probing level ϵ obtained by progressively higher density of targets ρ_T.

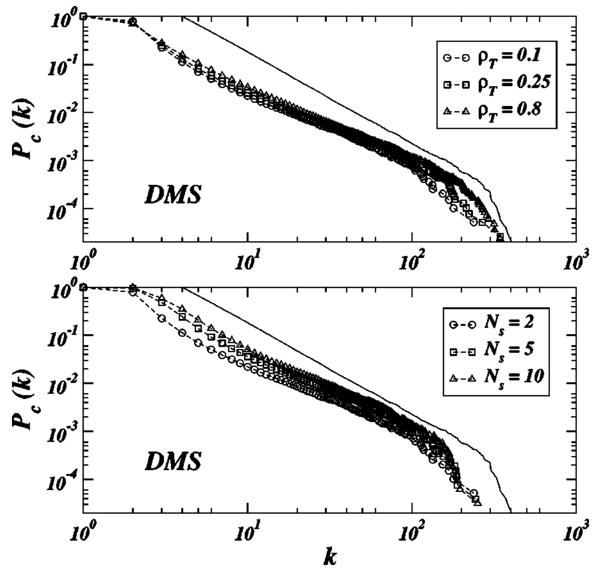

FIG. 6. Cumulative degree distribution of the sampled DMS graph for USP probes. The top figure shows sampled distributions obtained with $N_S=2$ and varying density target ρ_T. The figure on the bottom shows sampled distributions obtained with $\rho_T=0.1$ and varying number of sources N_S. The solid line is the degree distribution of the underlying graph.

ues of ϵ. The data show clearly that the low degree regime is instead undersampled, resulting in a bending of the curves or an apparent change in the exponent of the degree distribution, as already noticed by Petermann and De Los Rios in the case of single source mapping procedures [23].

According to the present analysis, 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 traceroutelike exploration.

V. CONCLUSIONS AND OUTLOOK

The presented statistical mean-field analysis of exploration techniques based on shortest-path routing provides a general interpretative framework for the results obtained in numerical experiments on graph models. The sampled graph clearly distinguishes the two situations defined by homogeneous and heterogeneous topologies, respectively. This is due to the exploration process which statistically focuses on high betweenness nodes, thus providing a very accurate sampling of distribution tails. Therefore, the main topological features of heavy-tailed networks are more easily discriminated, being the relevant statistical information primarily contained in the fairly well-captured degree distribution tail. The sampling of homogeneous graphs appears more cumbersome, but surprising effects such as the existence of apparent power laws are found only in very peculiar cases. According to our theoretical approach, multisource exploration procedures generally 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. Up to now available data indicate the presence of heavy-tailed degree distributions at both the router and AS levels. The upper degree cutoff at the router and AS level runs up to 10^2 and 10^3, respectively. Then, in the light of the present discussion, it is very unlikely that this feature is just an artifact of the mapping strategies. Indeed, a homogeneous graph should have an average degree comparable to the measured cutoff; this means that for the Internet to be a homogeneous graph it would require that nine routers out of ten would have more than 100 links to other routers, something quite unrealistic. In addition, the majority of mapping projects are multisource, a feature that we have shown to readily wash out the presence of spurious power-law behavior. On the contrary, power-law tails are easily sampled with enough accuracy for the large degree part at all probing levels. As a natural consequence, the heavy-tail behavior observed in real mapping experiments should be, very plausibly, a genuine feature of the Internet. On the other hand, it is very important to stress that, at the quantitative level, some properties, such as average degree, distribution exponent and clustering, might exhibit considerable deviations from their true values. In addition, degree correlation properties have been found in Internet maps that exhibit a disassortative cahracter [33,34]. This implies that large degree vertices tend to be connected to small degree ones, and vice versa for small degrees vertices; it would also be interesting to understand how such properties may affect the sampling. The models we used do not show any particular correlation structure and a preliminary numerical analysis does not appear to introduce spurious correlation effects in the sampled graph. Further tests on specific models with stronger correlations are beyond the scope of this paper but might provide interesting results in our understanding of the mapping process. In these respects, it is of major importance to define strategies in order to optimize the accuracy of the various parameters and quantities of the underlying graph.

In conclusion, in this paper we have proposed a statistical theory of the shortest-path probing of large information networks such as the Internet. We unveil, by means of a simple mean-field approximation, the relations between the statistical observables of the discovered graph and general topological properties of the unknown underlying network (such as the betweenness centrality). It is worth remarking that the property of centrality plays an important role in many dynamical processes occurring on networks, such as, e.g., epidemic spreading where the most central nodes are crucial in the propagation pattern. The relation between our capacity of measuring the structure of networks and the biases introduced by the vertices centrality may therefore be interesting in the forecast of computer virus epidemics and other digital attacks. Finally, we stress that the quantitative optimization of large network sampling is a more difficult and technical problem that calls for further detailed work aimed at a more precise assessment of the mapping strategies on both the analytic and numerical sides.

ACKNOWLEDGMENTS

This work has been partially supported by the European Commission Fet-Open project No. COSIN IST-2001-33555 and Contract No. 001907 (DELIS).

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[2] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[4] P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms* (Wiley, Chichester, 2003).

[5] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, U.K., 2004).

[6] The National Laboratory for Applied Network Research

(NLANR), sponsored by the National Science Foundation; see http://moat.nlanr.net/

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

[8] Topology project, Electric Engineering and Computer Science Department, University of Michigan; see http://topology.eecs.umich.edu/

[9] SCAN project at the Information Sciences Institute; http://www.isi.edu/div7/scan/

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

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).

[12] R. Govindan and H. Tangmunarunkit, in Proceedings of IEEE INFOCOM 2000, Tel-Aviv, Israel, March 2000, pp. 1371–1380.

[13] A. Broido and K. C. Claffy, San Diego Proceedings of SPIE International Symposium on Convergence of IT and Communication, Denver, Colorado, 2001.

[14] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).

[15] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001); A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).

[16] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, in Proceedings of IEEE INFOCOM 2002, New York, 2002.

[17] A. Medina and I. Matta, Boston University, Technical Report No. BU-CS-TR-2000-005, 2000 (unpublished).

[18] C. Jin, Q. Chen, and S. Jamin, EECS Department, University of Michigan, Technical Report No. CSE-TR-433-00, 2000 (unpublished).

[19] H. Burch and B. Cheswick, IEEE Trans. Comput. **32**, 97 (1999).

[20] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker, Proc. Natl. Acad. Sci. U.S.A. **99**, 2573 (2002).

[21] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie, Department of Computer Sciences, Boston University, Technical Report No. BUCS-TR-2002-021, 2002 (unpublished).

[22] A. Clauset and C. Moore, Phys. Rev. Lett. **94**, 018701 (2005).

[23] T. Petermann and P. De Los Rios, Eur. Phys. J. B **38**, 201 (2004).

[24] J.-L. Guillaume and M. Latapy, in Proceedings of IEEE INFOCOM 2005 (to be published).

[25] L. C. Freeman, Sociometry **40**, 35 (1977).

[26] U. Brandes, J. Math. Sociol. **25**, 163 (2001).

[27] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).

[28] P. Erdös and P. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).

[29] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[30] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[31] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **63**, 062101 (2001).

[32] M. Barthélemy, Eur. Phys. J. B **38**, 163 (2004).

[33] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[34] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).