What is the real size of a sampled network? The case of the Internet

Fabien Viger,[1] Alain Barrat,[2,3,4] Luca Dall'Asta,[2,3] Cun-Hui Zhang,[5] and Eric D. Kolaczyk[6]
[1]*LIP6, UMR 7606 du CNRS, Université de Paris-6, 4 place Jussieu, 75005, Paris, France*
[2]*LPT, UMR 8627 du CNRS, 91405 Orsay cedex, France*
[3]*Université Paris-Sud, 91405 Orsay cedex, France*
[4]*Complex Networks Lagrange Laboratory, ISI Foundation, Viale. S. Severo 65, 10133 Turin, Italy*
[5]*Department of Statistics, Rutgers University, 504 Hill Center, Busch Campus, Piscataway, NJ 08854–8019 USA*
[6]*Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215 USA*
(Received 10 January 2007; published 17 May 2007)

Most data concerning the topology of complex networks are the result of mapping projects which bear intrinsic limitations and cannot give access to complete, unbiased datasets. A particularly interesting case is represented by the physical Internet. Router-level Internet mapping projects generally consist of sampling the network from a limited set of sources by using traceroute probes. This methodology, akin to the merging of spanning trees from the different sources to a set of destinations, leads necessarily to a partial, incomplete map of the Internet. The determination of the real Internet topology characteristics from such sampled maps is therefore, in part, a problem of statistical inference. In this paper we present a twofold contribution in order to address this problem. First, we argue that inference of some of the standard topological quantities is, in fact, a version of the so-called "species" problem in statistics, which is important in categorizing the problem and providing some indication of its inherent difficulties. Second, we tackle the issue of estimating arguably the most basic of network characteristics—its number of nodes—and propose two estimators for this quantity, based on subsampling principles. Numerical simulations, as well as an experiment based on probing the Internet, suggest the feasibility of accounting for measurement bias in reporting Internet topology characteristics.

DOI: 10.1103/PhysRevE.75.056111 PACS number(s): 89.75.Hc, 89.20.Hh, 89.70.+c

I. INTRODUCTION

The enormous amount of work dedicated in the recent years to the study and understanding of complex networks [1–5] has been largely due to the possibility of accessing and analyzing unprecedented amounts of data. In particular, the interest of physicists has been stimulated by the observation of ubiquitous patterns such as (i) the small-world property [6], defined by an average shortest path length—average distance between any pair of vertices—increasing very slowly with the network size N; (ii) the presence of a large transitivity [7], which implies that two neighbors of a given vertex are also connected to each other with large probability; (iii) a scale-free behavior for the degree distribution $P(k)$, defined as the probability that a vertex is connected to k other vertices (has degree k), that typically shows power-law behavior $P(k) \sim k^{-\gamma}$, where γ is a characteristic degree exponent, usually in the range $2 < \gamma < 3$.

The data on which such observations are based are, however, often incomplete and the result of an incomplete sampling of the real network one would like to study. They may therefore *a priori* suffer from uncontrolled biases. Recently, the question of the accuracy of the topological characteristics inferred from such maps has been the subject of various studies, to understand in particular how various sampling techniques introduce biases that can alter the network's characteristics [8–16].

In this paper, we focus on the case of the physical Internet (by "physical Internet" we mean the network composed of routers—nodes—connected by cables and other communication systems through which the Internet traffic transits) to tackle the issue of how real characteristics of a network can be inferred from the sampled data, i.e., how the sampling biases can be corrected. Due to the lack of any centralized, complete map of the physical Internet, the measurement and characterization of Internet topology is indeed a task of fundamental interest. The most common approach to build and update partial Internet maps at this level consists in acquiring local views from several vantage points, using the well-known traceroute tool to evaluate paths to different destinations. The traceroute command sends probes (data packets) toward a certain Internet node (IP address) and provides the addresses of the traversed nodes. Various such projects have been developed in the last years and have allowed to obtain important information on the structure of the Internet [17–22], even if the corresponding maps are necessarily incomplete. The way in which traceroute measurements introduce sampling biases, as first observed in Ref. [8], has been analyzed, and the following consensus has been reached [9–14]: while qualitative conclusions on the topology drawn from traceroutelike sampling (e.g., a highly variable node degree distribution) are reliable, conclusions of a precise quantitative nature are much more subject to biases. That is, there is the possibility for considerable deviations between quantitative measurements of topological characteristics of the sampled Internet maps and those of the actual Internet.

The problem of accounting for such deviations can be seen, from a statistical perspective, as one of designing appropriate estimators of topological characteristics (e.g., average degree, clustering coefficients, etc.) that correct for the underlying bias. Designing estimators to match a given sampling design is a canonical task in statistical sampling theory (e.g., Ref. [23]). In fact, there exists a small but carefully

developed body of such work in the specific context of graph sampling, primarily in the social networks literature (e.g., see Ref. [24], and references therein). However, it may be observed from this work that the task in this context is particularly challenging, with estimators needing to incorporate aspects of the sampling design, effects of network topology, and the nature of the characteristic to be inferred. Furthermore, the solutions in this literature do not directly address the particular type of path-based sampling in traceroute studies.

Our contribution in this paper is to lay some initial groundwork on the topic of inferring Internet topology characteristics from traceroute-generated maps. In Sec. II we review notations and explain how the inference of some standard topological characteristics fall under the category of so-called "species" problems in statistics, a point which has fundamental implications. We then focus on the most basic of these Internet "species," namely, the number of nodes in a network (i.e., the network size). In Sec. III, we provide analytical arguments illustrating the difficulty of the problem, and propose two nonparametric estimators for the network size in Sec. IV. We present the results of a numerical simulation study and of a real Internet measurement experiment in Sec. V. These results suggest the feasibility of designing estimators that account for traceroute bias, although additional challenges are to be expected in estimating further graph characteristics, as discussed in Sec. VI.

II. BACKGROUND

A. Model and notation

Throughout this paper we will represent an arbitrary network of interest as an undirected, connected graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is a set of vertices (nodes) and $\mathcal{E}$ is a set of edges (links). We denote by $N=|\mathcal{V}|$ and $M=|\mathcal{E}|$ the numbers of vertices and edges, respectively. As a model for a typical traceroute study, we consider that a set $S=\{s_1,\ldots,s_{n_S}\}$ of n_S active sources deployed in the network sends probes to a set $T=\{t_1,\ldots,t_{n_T}\}$ of n_T destinations (or targets), with $S,T\subset\mathcal{V}$. We also define the source and destination densities $q_S=n_S/N$ and $q_T=n_T/N$.

Each probe collects information on all the vertices and edges traversed along the path connecting a source to a destination [25]. The merging of the various sampled paths yields a partial map of the network, that may in turn be represented as a sampled subgraph $\mathcal{G}^*=(\mathcal{V}^*,\mathcal{E}^*)$. The actual paths followed by the probes depend on many different factors, such as commercial agreements, traffic congestion, and administrative routing policies. In fact, the issue of modeling the paths from source to destination is in itself nontrivial. It is therefore important to emphasize that the theoretical work and the estimators presented in the next sections are independent on the specific routing model used. In the numerical validation (Sec. V) of our theoretically derived estimators, however, we will use a very simple model in which probes use shortest paths between sources and targets. Numerical simulations with more refined models such as the one presented in Ref. [26] do give similar results.

It should be noted, however, that in our framework we do not take into account the various anomalies that can arise in traceroute-based measurement studies. These anomalies may introduce complex artifacts in IP topology maps constructed from traceroute data, as described in Ref. [27], and are beyond our scope. The issue of nonresponding routers (resulting in incomplete paths) is similarly not discussed in this paper either. We will instead assume that the traceroute tool provides a realistic, unique, and complete path from a source to a destination. In principle, our framework can be expanded to account for issues such as those raised above, but at the cost of additional analytic and computational complexity. We leave such extensions to future work, and concentrate here simply on developing basic foundations and principles.

B. Network inference as "species" problem

Let us consider a summary characteristic of the topology of $\mathcal{G}$, which we denote by $\eta(\mathcal{G})$, such as the number of vertices N, the number of edges M, or the collection of node degrees $\{k_i\}_{i\in\mathcal{V}}$. The observed values $N^*=|\mathcal{V}^*|$ and $M^*=|\mathcal{E}^*|$, as simple totals, necessarily underestimate N and M unless sampling is exhaustive. In fact, the studies on traceroutelike sampling of networks [8–10,12–14] have shown that the observed values can differ strongly from the real values, with consequently very important sampling biases. In fact, typical traceroute studies explore much more thoroughly the vertices and edges near sources and targets, as well as "central" vertices through which many paths go, and ignore many low-degree vertices [11,12].

Given the existence of the strong quantitative biases in the measured characteristics of sampled networks, and in particular of the Internet, the question naturally arises as to whether or not it is possible to produce more accurate estimates of topology characteristics from traceroute samples. An important initial step towards answering this question is the observation that estimation of the quantities N, M, and $\{k_i\}$ in this context falls under the category of so-called "species" problems in statistics. Stated generically, the species problem refers to the situation in which, having observed a certain number of members of a population, each of whom falls into one of C distinct classes (or "species"), an estimate $\hat{C}$ of C is desired. A typical example is that of the proverbial biologist sitting in a forest and observing animals. Although at the end of the day the biologist may have observed, say, four lions, two tigers, and six bears, the total number C^* of species observed is only three, and the biologist would like to know how many other species are present but were not observed. This problem arises in numerous contexts and has received a lot of attention in statistics (see Ref. [28] for an overview and an extensive bibliography). Perhaps surprisingly, however, while the estimation of the relative frequencies of species in a population is straightforward (given knowledge of the total number of species C), the estimation of C itself is often difficult. The potential difficulty is due to the fact that species present in relatively low proportions in the population are expected to be missed, and there could be an arbitrarily large number of such species in arbitrarily low proportions.

The estimation of N, M, and degrees $\{k_i\}$ can be rephrased as species problems with traceroute sampling. For example, in estimating N, consider each separate vertex i as a "species" and declare that the species i has been observed each time i is encountered on one of the $n=n_S n_T$ traceroute paths. In other words, a vertex will represent either a "common" species if it lies on many of the collected paths, or a "rare" species if it was observed only once or not at all. With our notations, the total number of species is N and the observed number of species is N^*. A similar argument shows that estimation of the number of edges M too may be mapped to a species problem. Finally, as argued in Ref. [29], the problem of inferring the degree k_i of a vertex i from traceroute measurements can also be mapped to a species problem, by letting all edges incident to i constitute a species and declaring this species to have been observed every time one of those edges is encountered.

As a first step in this direction, we focus on the task of estimating N, the total number of vertices in the network graph $\mathcal{G}$. In the case of the Internet, it will correspond to the number of "alive" IP addresses in the network. Focusing on N is a logical choice in order to assess the feasibility of our approach since it should correspond to the simplest quantity to estimate: we do not expect to be able to produce estimators for M, k_i, or more complex metrics if we fail on N first. This constitutes therefore a necessary and nontrivial first step, as shown in the following sections.

III. INFERRING N: CHARACTERIZATION OF THE PROBLEM

Before proceeding to the construction of estimators for N, as we will do in Sec. IV, it is useful to first better understand the structural elements of the problem. In particular, the following analysis provides insight into the structure of the underlying "population," the relative frequency of the various "species," and the impact of these factors on the problem of inferring N. For the sake of exposition, and as discussed in the previous section, we adopt here the convention of modeling Internet traceroute routing, to a first approximation, as "shortest-path" routing. The discussion could, however, be extended to the case of other routing models.

A. The betweenness centrality

A crucial quantity in the characterization of traceroutelike sampling is the so-called betweenness centrality, which essentially counts for each vertex the number of shortest paths on which it lies: nodes with large betweenness lie on many shortest paths and are thus more easily and more frequently probed [11,12]. More precisely, if $\mathcal{D}_{hj}$ is the total number of shortest paths from vertex h to vertex j, and $\mathcal{D}_{hj}(i)$ is the number of these shortest paths that pass through the vertex i, the betweenness of i is defined as $b_i=\Sigma\mathcal{D}_{hj}(i)/\mathcal{D}_{hj}$, where the sum runs over all h,j pairs with $j\neq h\neq i$ [30,31]. It can be shown [32] that the average shortest path length between pairs of vertices l is related to the betweenness centralities through the expression

$$\sum_i b_i = N(N-1)(l-1).$$

This may be rewritten in the form

$$N = 1 + \frac{E[b]}{l-1}, \tag{1}$$

where $E[b]$ denotes the average betweenness centrality of the nodes.

Empirical experiments suggest that the average shortest path length l can be estimated quite accurately, which is not surprising given the path-based nature of traceroute. The problem of estimating N is thus essentially equivalent to that of estimating the average betweenness centrality $E[b]$.

It turns out that many real world networks, and in particular Internet maps, have been found to display broad distributions of betweenness centrality values, with an approximate asymptotical power-law shape [4]. Moreover our numerical investigations—in the same spirit as Ref. [11]—show that this power-law shape is robust with respect to traceroutelike sampling, meaning that the real exponent of the asymptotic power law can reasonably be estimated by the measured exponent.

Let us first emphasize that the betweenness distribution is not expected to follow exactly a power law on the whole distribution range. Instead, it is more realistic to picture the distribution $P(b)$ as a mixture distribution [33]:

$$P(b) = \pi P_1(b) + (1-\pi)P_2(b), \tag{2}$$

where P_1 is a distribution at low values $b\in[1,b_{\min})$, for some $b_{\min}$ small, and $P_2(b)$ is a distribution at large values $b\in[b_{\min},b_{\max}]$, $b_{\max}\gg b_{\min}$ that follows a power law $P_2(b)=b^{-\beta}/K$. We note that this functional form is certainly not exact and already contains some arbitrariness.

B. Using the betweenness centrality to estimate N

We now consider the task of estimating N in the case that an ansatz similar to that in Eq. (2) holds true. As remarked, this problem is equivalent to that of estimating $E[b]$. Under Eq. (2), the average $E[b]$ in Eq. (1) is a weighted combination of two terms, i.e., $E[b]=\pi E_1[b]+(1-\pi)E_2[b]$. From the perspective of the simple ansatz just described, the challenge of accurately estimating $E[b]$—and hence N—can be viewed as a problem of the accurate estimation of the two means $E_1[b]$ and $E_2[b]$ and the weight π. The estimation of the first part $E_1[b]$ requires knowledge of the betweenness of vertices with "small" betweenness, i.e., of nodes $i\in\mathcal{V}$ traversed by relatively few paths. These are, however, precisely the nodes on which we receive the least information from traceroutelike studies, as they are expected to be visited infrequently or not at all. The relative proportion π of such nodes seems moreover to be similarly difficult to determine. As mentioned earlier, this is a hallmark characteristic of the species problem, i.e., the lack of accurate knowledge of the relative number in the population of species observed comparatively less frequently.

For the second mean E_2, we obtain, since $K=\int_{b_{\min}}^{b_{\max}} b^{-\beta}db$:

$$E_2[b] = \frac{1}{K}\int_{b_{\min}}^{b_{\max}} b^{1-\beta}db = \frac{b_{\min}^{2-\beta} - b_{\max}^{2-\beta}}{b_{\min}^{1-\beta} - b_{\max}^{1-\beta}}\frac{\beta-1}{\beta-2}.$$

This leads to

$$E_2[b] = \frac{\beta-1}{\beta-2}b_{\min}\frac{1-\delta^{\beta-2}}{1-\delta^{\beta-1}}, \quad \text{where } \delta = \frac{b_{\min}}{b_{\max}}.$$

Additionally, if the only origin of the cutoff $b_{\max}$ is the finite size of the network, $b_{\max}$ can be defined by imposing the condition that the expected number of nodes beyond the cutoff is bounded by a fixed constant [3]

$$N \times \int_{b_{\max}}^{\infty} P(b)db \sim 1 \Rightarrow b_{\max} \sim \left(\frac{(\beta-1)K}{(1-\pi)N}\right)^{1/(1-\beta)}. \quad (3)$$

Therefore, assuming that $b_{\min} \ll b_{\max}$, and thus $K \sim b_{\min}^{1-\beta}/(\beta-1)$, we obtain

$$b_{\max} \sim b_{\min}[(1-\pi)N]^{1/(\beta-1)} \quad (4)$$

which gives for δ the approached value

$$\delta \sim [(1-\pi)N]^{1/(1-\beta)}$$

and, assuming that $1 \ll (1-\pi)N$, we finally obtain

$$E_2[b] \sim \frac{\beta-1}{\beta-2}b_{\min}(1-[(1-\pi)N]^{-(\beta-2)/(\beta-1)}). \quad (5)$$

$E_2[b]$ is therefore strongly dependent on the ratio $\frac{\beta-1}{\beta-2}$. It turns out, however, that the empirical values of β in the actual Internet maps collected so far are very close to 2. (Similarly, values close to 2 are obtained as well in many models of heterogeneous networks [34,35].) This implies a highly unstable estimation of $E_2[b]$ with respect to the measurement uncertainty in the value of β.

In summary, while it appears that both the average path length l and the power-law shape parameter β can be estimated in a fairly stable fashion from traceroutelike data in and of themselves, this is likely insufficient to allow us to obtain from them an accurate estimate of N. First, because the uncertainties in estimates of β will be magnified in the estimate of $E_2[b]$, and second because the data can be expected to have little information for directly estimating π and $E_1[b]$. Moreover, the functional form assumed for $P(b)$ itself bears some arbitrariness, which introduces still more uncertainty.

The above analysis both highlights the relevant aspects of the species problem inherent in estimating N and indicates the substantial difficulties of a classical parametric estimation approach. One is led, therefore, to consider parameter-free methods, such as those we develop in the next sections.

IV. ESTIMATION OF NETWORK SIZE

Starting from the knowledge of the observed number of nodes N^* in $\mathcal{G}^*$, we propose two estimators, both of which essentially have the form $\hat{N} \approx N^*/\alpha$, where $\alpha \in (0,1)$ is a data-dependent factor that inflates N^*. The specific nature of this factor in each case derives from a formal argument based on statistical subsampling principles.

A. A resampling estimator

A popular subsampling method consists in resampling, which underlies the well-known "bootstrap" method [36]. Given a sample $X^* = \{x_1^*, \ldots, x_m^*\}$ from a population $X = \{x_1, \ldots, x_n\}$, resampling in its simplest form means taking a second sample $X^{**} = \{x_1^{**}, \ldots, x_l^{**}\}$ from X^* to study a certain relationship between the first sample X^* and the true population X through the *observed* relationship between the second and first samples X^{**} and X^*. We use a similar principle here, through which the relation between characteristics of $\mathcal{G}^*$ and $\mathcal{G}$ is inferred from the relation between a sample $\mathcal{G}^{**}$ of $\mathcal{G}^*$ and $\mathcal{G}^*$ itself.

Let us denote by n_S^* and n_T^* the number of sources and destinations used for the resampling performed on $\mathcal{G}^*$. The corresponding densities are denoted by $q_S^* = n_S^*/N^*$ and $q_T^* = n_T^*/N^*$. Now, consider the quantity N^*/N, i.e., the fraction of nodes discovered through traceroute sampling of $\mathcal{G}$, which we will call the "discovery ratio." The expected discovery ratio $E[N^*/N]$ has been found to vary smoothly as a function of the fraction $q_T = n_T/N$ of targets sampled, for a given number n_S of sources [12,14]. Our resampling-based estimator is based on the assumption that the sampled subgraph $\mathcal{G}^*$ is sufficiently representative of $\mathcal{G}$ so that a sampling on $\mathcal{G}^*$ similar to that used in its obtention from $\mathcal{G}$ yields a discovery ratio similar to the fraction of nodes discovered in $\mathcal{G}$. Formally, it corresponds to the following property:

$$\left.\begin{aligned} q_S &= q_S^* \\ q_T &= q_T^* \end{aligned}\right\} \Rightarrow \frac{E[N^{**}]}{N^*} = \frac{N^*}{N}, \quad (6)$$

where the expectation $E[N^{**}]$ is with respect to whatever random mechanism drives the choice of source and target sets S^* and T^* on $\mathcal{G}^*$.

The condition of equal discovery rates can be rewritten in the form $N \sim N^*(N^*/E[N^{**}])$. The quantity $E[N^{**}]$ can be estimated by repeating the resampling experiment a certain number B of times, compiling subgraphs $\mathcal{G}_1^{**}, \ldots, \mathcal{G}_B^{**}$ of sizes $N_1^{**}, \ldots, N_B^{**}$, and forming the average $\bar{N}^{**} = (1/B)\Sigma_k N_k^{**}$. Substitution then yields

$$\hat{N}_{\mathrm{RS}} = N^*\frac{N^*}{\bar{N}^{**}} \quad (7)$$

as a resampling-based estimator for N.

Note, however, that its derivation is based upon the premise that $q_S^* = q_S$ and $q_T^* = q_T$, and q_S, q_T are in fact unknown (i.e., since N is unknown). This issue is addressed in the following way: we first replace the hypothesis $q_S^* = q_S$ by $n_S^* = n_S$, since typically the number of sources is very small and n_S is a more relevant quantity than q_S [11,12]. We have moreover performed numerical studies on networks with various topological characteristics, namely, a Barabási-Albert (BA) network and an Erdös-Renyi (ER) network, which are prototypical examples of heterogeneous and homogeneous networks, respectively. We have also used a real Internet map obtained from the Skitter project (see Sec. V for more details). These empirical studies, using uniform random sampling of source and target nodes, suggest that the modified assumption

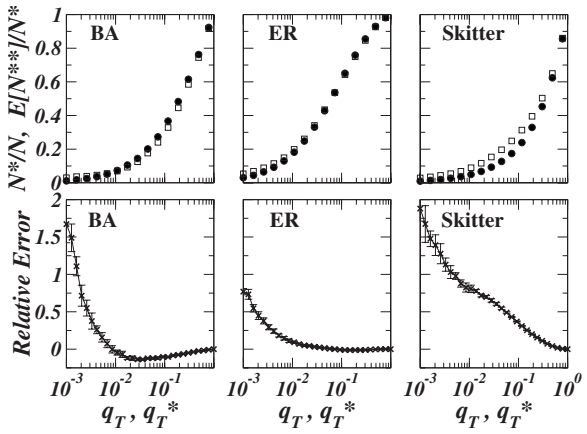

FIG. 1. A comparison of the quantities N^*/N and $E[N^{**}]/N^*$, as a function, respectively, of $q_T=n_T/N$ and $q_T^*=n_T^*/N^*$, for the three networks described in Sec. V. Here $n_S=n_S^*=10$. The top row shows the averages of N^*/N and $E[N^{**}]/N^*$ over ten realizations of $\mathcal{G}^*$. The bottom row shows the average of the difference of these two quantities, relative to N^*/N, over the same ten realizations. The comparison in the top row confirms the validity of the assumption (8) underlying the resampling estimator derived in Sec. IV A, while the comparison in the bottom row indicates that better performance of the estimator can be expected with increasing q_T (one standard deviation error bars are smaller than the symbol size in most cases).

$$\left.\begin{aligned} n_S &= n_S^* \\ q_T &= q_T^* \end{aligned}\right\} \Rightarrow \frac{E[N^{**}]}{N^*} = \frac{N^*}{N}, \tag{8}$$

holds reasonably well over a broad range of sampling ratios, as shown in Fig. 1.

The equation $q_T=q_T^*$ is moreover rewritten in the equivalent form $\frac{n_T^*}{n_T}=\frac{N^*}{N}$. The hypothesis (8) then implies the following identity, which involves only measurables or known quantities

$$\left.\begin{aligned} n_S &= n_S^* \\ q_T &= q_T^* \end{aligned}\right\} \Rightarrow \frac{E[N^{**}]}{N^*} = \frac{n_T^*}{n_T}. \tag{9}$$

As depicted in Fig. 2 for the BA network, this suggests a simple, dichotomic method to adjust n_T^* until the relation (9) holds, by searching the intersection point of the curves $(x=n_T^*/n_T, y=\bar{N}^{**}/N^*)$ and $(y=x)$. The value of $\bar{N}^{**}$ for the appropriate n_T^* is then substituted into Eq. (7) to produce $\hat{N}_{\mathrm{RS}}$. In practice, one may either use a fixed value of B (recall that B denotes the number of times the resampling is performed to get an estimation of $E[N^{**}]$) throughout or, as we have done, increase B as the algorithm approaches the condition $n_T^*/n_T \approx \bar{N}^{**}/N^*$.

B. A "leave-one-out" estimator

Another popular subsampling paradigm is the "leave-one-out" strategy underlying such methods as jack knifing and cross-validation [36]. The same underlying principle may be applied in a useful manner here to the problem of estimating N, in a way that does not require the assumptions underlying Eq. (7), as we now describe.

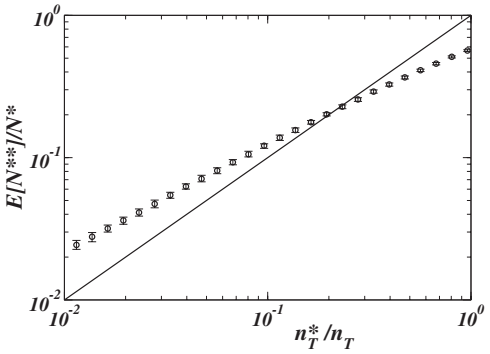

FIG. 2. Illustration of the obtention of the resampling estimator, in the case of a BA graph (see Sec. V for more details on the networks) of size $N=10^5$. The initial sampling was obtained with $n_S=10$ sources and $n_T=10^4$ targets ($q_T=0.1$), yielding a graph $\mathcal{G}^*$ of size $N^*=33\,178$. The circles show the ratio of the average size of the resampled graph $\mathcal{G}^{**}$, $\bar{N}^{**}/N^*$, as a function of the ratio n_T^*/n_T, with $n_S^*=n_S=10$ sources. The error bars give the variance with respect to the various placements of sources and targets used for the resampling. The straight line is $y=x$ and allows one to find the value of n_T^* such that $n_T/n_T^*=N^*/\bar{N}^{**}$.

Recall that $\mathcal{V}^*$ is the set of all vertices discovered through a traceroute study, including the n_S sources $S=\{s_1,\ldots,s_{n_S}\}$ and the n_T targets $T=\{t_1,\ldots,t_{n_T}\}$. Our approach is to connect N to the frequency with which individual targets t_j are included in traces from the sources in S to the other targets in $T\backslash\{t_j\}$. Accordingly, let $\mathcal{V}_{(-j)}^*$ denote the number of vertices discovered by traces to targets other than t_j, and define $\delta_j=I\{t_j\notin\mathcal{V}_{(-j)}^*\}$ to be the indicator of the event that target t_j is not "discovered" by traces to any other target (i.e., $\delta_j=1$ if t_j is not discovered by traces to the other targets and $\delta_j=0$, otherwise). The total number of such targets can be written as $X=\Sigma_j\delta_j$. The basic idea of the estimator is to derive a relation between X and N. The measure of X during a sampling experiment will then allow one to estimate N.

We assume that, given a preselected set of source nodes (chosen either randomly or not), the set of targets is chosen at random from the remaining vertices in V. The probability that target t_j is not discovered by the paths to other targets is then simply given by

$$\Pr(\delta_j=1|\mathcal{V}_{(-j)}^*)=\frac{N-N_{(-j)}^*}{N-n_S-n_T+1}, \tag{10}$$

where $N_{(-j)}^*=|\mathcal{V}_{(-j)}^*|$. Note that, by symmetry, the expectation $E[N_{(-j)}^*]$ is the same for all j: we denote this quantity by $E[N_{(-)}^*]$. As a result, we obtain

$$E[X]=\sum_{j=1}^{n_T}\Pr(\delta_j=1|\mathcal{V}_{(-j)}^*)=\frac{n_T(N-E[N_{(-)}^*])}{N-n_S-n_T+1}. \tag{11}$$

Rewriting this equation to isolate N, we have

$$N=\frac{n_TE[N_{(-)}^*]-(n_S+n_T-1)E[X]}{n_T-E[X]}. \tag{12}$$

An estimator for N may be obtained from the previous expression (12) by estimating the unknown components on the right-hand side, namely, $E[N^*_{(-)}]$ and $E[X]$. It seems natural to use the unbiased estimators $\bar{N}^*_{(-)}=(1/n_T)\Sigma_j N^*_{(-j)}$ and X itself, which is measured during the traceroute study. However, while substitution of these quantities in the numerator of Eq. (12) is straightforward, substitution of X for $E[X]$ in the denominator can be problematic in the case that $X=n_T$. Indeed, when none of the targets t_j are discovered by traces to other targets, as is possible if $q_T=n_T/N$ is small, N will be estimated by infinity. A better strategy is to estimate the quantity $1/(n_T-X)$ directly. This produces the estimator (see Ref. [37] for the technical details, as well as an estimate of the variance of this estimator)

$$\hat{N}_{L1O}=\frac{n_T+1}{n_T}\cdot\frac{n_T\bar{N}^*_{(-)}-(n_S+n_T-1)X}{n_T+1-X}. \tag{13}$$

The primary assumption underlying this derivation is the condition that $N^*_{(-j)}\approx N^*_{(-j')}\approx N^*_{(-j,-j')}$, where $N^*_{(-j,-j')}=|\mathcal{V}^*_{(-j)}\cap\mathcal{V}^*_{(-j')}|$. This condition is well motivated by empirical findings in the literature (as well as our own numerical experiments), in that it is equivalent to saying that the unique contribution of discovered vertices by traces to any one or any pair of target vertices is relatively small. For example, using data collected by the Skitter project at CAIDA [18], a fairly uniform discovery rate of roughly three new nodes per new target, after the initial 200 targets, has been cited [38].

Note that this condition also implies that $N^*_{(-j)}\approx N^*$, for all j, which suggests replacement of $\bar{N}^*_{(-)}$ by N^* in Eq. (13). Upon doing so, and after a bit of algebra, we arrive at the approximation

$$\hat{N}_{L1O}\approx(n_S+n_T)+\frac{N^*-(n_S+n_T)}{1-w^*}, \tag{14}$$

where $w^*=X/(n_T+1)$. In other words, $\hat{N}_{L1O}$ can be seen as counting the n_S+n_T vertices in $S\cup T$ separately, and then taking the remaining $N^*-(n_S+n_T)$ nodes that were discovered by traces and adjusting that number upward by a factor of $(1-w^*)^{-1}$. This form is in fact analogous to that of a classical method in the literature on species problems, due to Good [39], in which the observed number of species is adjusted upwards by a similar factor that attempts to estimate the proportion of the overall population for which no members of species were observed. Such estimators are typically referred to as coverage-based estimators, and a combination of theoretical and numerical evidence seems to suggest that they enjoy somewhat more success than most alternatives [28].

V. NUMERICAL RESULTS

A. Simulation study

1. Methodology

We examined the performance of the proposed estimators using a methodology similar to those developed in Refs. [9,12,14]. The idea is to start from known graphs $\mathcal{G}$ of given size N, having various topological characteristics, equipped each with an assumed routing structure. For each graph $\mathcal{G}$, a traceroutelike sampling is performed, yielding a sampled graph $\mathcal{G}^*$. The estimators $\hat{N}_{\mathrm{RS}}$ and $\hat{N}_{L1O}$ are then computed and compared with both the size of the sampled graph $\mathcal{G}^*$ and the original size N. The process is repeated a number of times, for various choices of source and target nodes, and for different values of the sampling effort, i.e., of the numbers of sources and targets n_S and n_T, and for various values of the initial size N. A performance comparison of the various estimators is then made by comparing values of $\hat{N}/N$, for $\hat{N}=N^*$, $\hat{N}_{\mathrm{RS}}$, and $\hat{N}_{L1O}$.

We present here the results obtained on three network topologies, two synthetic and one based on measurements of the real Internet. The synthetic topologies were generated according to (i) the classical Erdös-Rényi (ER) model and (ii) the Barabási-Albert (BA) model of scale-free networks. These models yield indeed the simplest and most well-known examples of graphs with homogeneous and heterogeneous degree distributions, respectively, and allow us therefore to test the proposed estimators on networks with very different topological characteristics. None of these models is intended to give a faithful representation of the Internet, and we therefore use them with illustrative purposes. However, as the ER and BA topologies lack important characteristics of the real Internet, such as clustering, complex hierarchies, etc., as a third network we used an Internet map from the Skitter project [40], which consisted in a traceroute sample taken in May 2006 from 18 sources sending probes towards 445 768 destinations, all around the world [41]. These three topologies were chosen as best representatives from a larger set of topologies that we actually used for our experiments, which were either synthetic (variations of the Barabási-Albert model, random graphs with power-law degree distributions) or coming from actual Internet maps (MERCATOR [42] Internet map from 1999, and CAIDA [18] Internet map from 2003), and which yielded very similar results to the ones presented here.

More precisely, we have used randomly generated ER and BA networks with average degree 6, and sizes N ranging from 10^3 to 10^6 nodes. The Internet sample from Skitter yielded a graph with $N=624\,324$ nodes and $M=1\,191\,525$ edges. Given a graph $\mathcal{G}$, and a chosen set of values for N, n_S, and n_T, a traceroutelike study was simulated as follows. First, a set of n_S sources $S=\{s_1,\ldots,s_{n_S}\}$ were sampled uniformly at random from $\mathcal{V}$ and a set of n_T targets $T=\{t_1,\ldots,t_{n_T}\}$ were sampled uniformly at random from $\mathcal{V}\backslash S$. Paths from each source to all targets were then extracted from $\mathcal{G}$, and the merge of these paths was returned as $\mathcal{G}^*$. Shortest path routing was used in collecting these simulated traceroutelike data, based on standard algorithms [43]. Unique shortest paths were forced by breaking ties randomly. After initial determination, routes are considered fixed, in that the route between a source $i\in S$ and a vertex $v\in V$ is always the same, independent of the destination target $j\in T$.

2. Simulation results

The plots in Fig. 3 show a comparison of N^*/N, $\hat{N}_{\mathrm{RS}}/N$, and $\hat{N}_{L1O}/N$, for $n_S=1$, 10, and 100 sources, as a function of

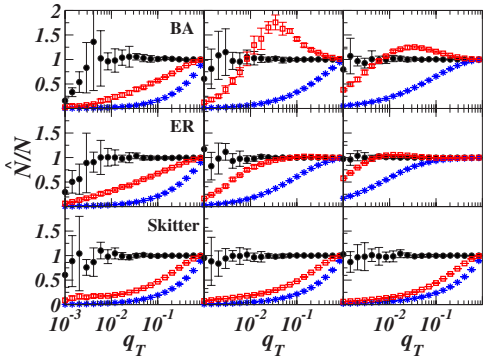

FIG. 3. (Color online) Comparison of the various estimators for the BA (top), ER (middle), and Skitter (bottom) networks. The curves show the ratios of the various estimators to the true network size, as a function of the target density q_T. Full circles: $\hat{N}_{L1O}/N$; empty squares: $\hat{N}_{\mathrm{RS}}/N$; stars: N^*/N. Values and one standard deviation error bars are based on 100 trials, with random choice of sources and targets for each trial. Left figures: $n_S=1$ source; middle: $n_S=10$ sources; right: $n_S=100$ sources.

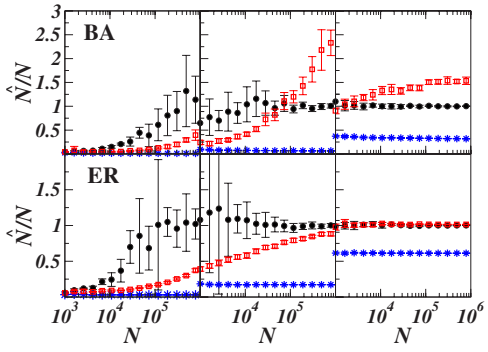

FIG. 4. (Color online) Effect of the size N of the graph $\mathcal{G}$ for BA and ER graphs at constant number of sources and density of targets. The curves show the ratios of the various estimators to the true network size, as a function of the graph size N. Full circles: $\hat{N}_{L1O}/N$; empty squares: $\hat{N}_{\mathrm{RS}}/N$; stars: N^*/N. Values and one standard deviation error bars are based on 100 trials, with random choice of sources and targets for each trial. $n_S=10$. Left figures: $q_T=10^{-3}$; middle: $q_T=10^{-2}$; right: $q_T=10^{-1}$.

q_T. We note that $n_S=1$ is a very pessimistic case that, however, allows us to test the performance of the estimators in extreme cases. Values of 10 or 100 are more realistic, while the most recently launched Internet mapping projects aim at the use of thousands of sources [22]. A value of 1 for these ratios is desired, and it is clear that in the case of both the resampling and the "leave-one-out" estimator that the improvement over the trivial estimator N^* is substantial. Increasing either the number of sources n_S or the density of targets q_T yields better results, even for N^*, but the estimators we propose converge much faster than N^* towards values close to the true size N.

Between the resampling and the "leave-one-out" estimator, the latter appears to perform much better. For example, we note that while both estimators suffer from a downward bias for very low values of q_T, this bias persists into the moderate and, in some cases, even high range for the resampling estimator. This is probably due to the fact that the basic assumptions underlying the derivation of $\hat{N}_{\mathrm{RS}}$ are only approximately satisfied, while for $\hat{N}_{L1O}$, the underlying hypotheses are indeed well satisfied. Notice, however, that the "leave-one-out" estimator has a larger variability at small values of q_T, while that of the resampling estimator is fairly constant throughout. This is because the same number B of resamples is always used in calculating $\hat{N}_{\mathrm{RS}}$ in Eq. (7), for all q_T, and the uncertainty can be expected to scale with B, but in calculating $\hat{N}_{L1O}$ in Eq. (13), the uncertainty will depend on n_T (and hence on q_T).

In terms of topology, estimation of N appears to be easiest for the ER model. Even N^* is more accurate, i.e., the discovery ratio is larger. Estimation on the Skitter graph appears to be the hardest, likely because the Skitter graph has a much higher proportion of low-degree vertices than the two synthetic graphs, which therefore lie on very few paths and are very difficult to discover. Interestingly, however, the performance of the "leave-one-out" estimator seems to be quite stable in all three graphs. On a side note, we mention that the resampling estimator behaves in a rather curious, nonmonotonic fashion in two of the plots, as q_T grows. At the moment, we do not have a reasonable explanation for this behavior, although we note that it appears to be limited to the case of the BA graph.

In Fig. 4, we investigate, at fixed n_S and q_T, the effect of the real size of the graph N. The estimators perform better for larger sizes, while N^*/N on the contrary decreases. This is due to the fact that the sample graph $\mathcal{G}^*$ gets larger, providing more and richer information, even if the discovery ratio does not grow. The odd nature of the results for the BA graph comes from the peak associated with the resampling estimator mentioned earlier; see Fig. 3.

We have also considered the case of fixed numbers of sources and targets, for increasing size N; such a scenario would be faced if more and more Internet mapping efforts were not deployed with a growing Internet network. As could intuitively be expected, the quality of the estimators $\hat{N}_{\mathrm{RS}}$ and $\hat{N}_{L1O}$ then gets worse as N increases, as shown in Fig. 5, but $\hat{N}_{L1O}$ still performs remarkably well.

B. A small internet experiment

Our long-term interest is in the reliable estimation of arbitrary router-level topology characteristics from traceroute data. The case of estimating N has been studied above primarily as an important first step. However, if estimation of N alone were the only goal, there are natural alternatives that one might consider, and these could provide us with useful sources of additional evaluation. For example, an experiment could use ping, a command that is able to test the reachability of any IP address: testing the response of some sufficient number n of randomly chosen IP addresses could yield an estimator $\hat{a}$ of the fraction of "alive" addresses and, in turn, an estimator $\hat{N}_{\mathrm{ping}}=2^{32}\hat{a}$ that is much simpler than either of those proposed in this paper.

We have performed such an experiment on the Internet. A total of $n=3\,726\,773$ ping were sent from a single source,

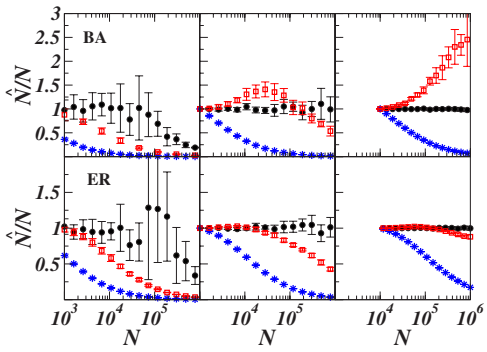

FIG. 5. (Color online) Effect of the size N of the graph $\mathcal{G}$ for BA and ER graphs at constant number of sources and targets. The curves show the ratios of the various estimators to the true network size, as a function of the graph size N. Full circles, $\hat{N}_{L1O}/N$; empty squares, $\hat{N}_{\mathrm{RS}}/N$; stars, N^*/N. Values and one standard deviation error bars are based on 100 trials, with random choice of sources and targets for each trial. $n_S=10$. Left figures, $n_T=10^2$ targets; middle, $n_T=10^3$ targets; right, $n_T=10^4$ targets.

yielding 61 246 valid responses (for a 1.64% response rate), and resulting in an estimate $\hat{N}_{\mathrm{ping}}=70\,583\,737$. We then performed a traceroute study from the same source to the 61 216 unique IP addresses obtained from the ping experiment, and calculated a "leave-one-out" estimate on the resulting $\mathcal{G}^*$ of $\hat{N}_{L1O}=72\,296\,221$. Of course, neither of these numbers are intended to be taken too seriously in and of themselves. The point is that, while the estimator from traceroute data is arguably less intuitive and direct in its derivation than that from the ping data, for the particular task of estimating N, it nonetheless produces roughly the same number. And, most importantly, while the ping data would of course not be useful for estimating M or degree characteristics, for example, the use of traceroute measurements, which produce an entire sampled subgraph $\mathcal{G}^*$, does in principle allow for the estimation of either of these quantities.

VI. CONCLUSIONS

In this paper, we have investigated the problem of correcting the inherent sampling biases of path-based samplings, which are commonly used to obtain partial maps of complex networks, and in particular of the Internet. We have shown how to recast this problem in the framework of the so-called "species" problem and, as a first illustrative case, we have focused on the issue of estimating the number N of nodes in the network. We have derived two different estimators based on subsampling principles. These estimators have then been tested numerically on networks with different topological characteristics, equipped with a simple model of traceroute-like routing. The numerical results have clearly shown the feasibility and interest of such approaches. As could be expected, the quality of the estimators increases with the initial

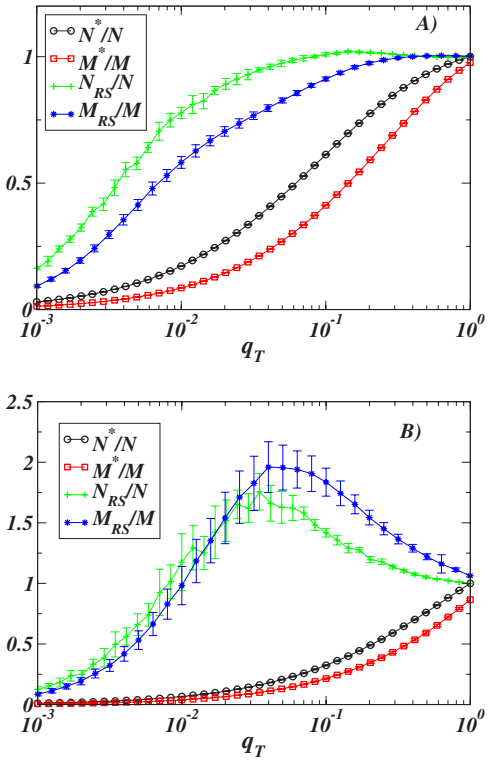

FIG. 6. (Color online) Performance of resampling estimators for the number of edges M_{RS} compared with the performance of $\hat{N}_{\mathrm{RS}}$. (A) ER network; (B) BA network. In both cases $N=10^5$, average degree 6, and $n_S=10$.

sampling effort, i.e., with the number of sources and density of targets of the traceroute sampling. In real Internet mapping experiments, this density is, of course, unknown; repeating the computation of the estimators for various experiments with increasing probing efforts should, however, allow us to obtain more and more reliable results. Moreover, while the number of sources has traditionally been quite small, recently proposed Internet mapping initiatives appear to be moving away from that trend (e.g., Ref. [22]). Our results suggest that, with larger numbers of sources, quite accurate estimates of N may be obtained even at very low levels of target density.

Future work will need to address the estimation of other network characteristics, such as M or degrees k_i. In this regard, while our results showed that the "leave-one-out" estimator performed noticeably better than the resampling estimator for estimating N, nevertheless the resampling estimator should not be summarily dismissed. While the derivation of the "leave-one-out" estimator is quite specific to the problem of estimating N, the derivation of the resampling estimator is general and independent of what is to be estimated. For example, it is straightforward to specify conditions for M analogous to those specified for N in Eq. (8), and the results of initial experiments shown in Fig. 6 indicate that a resampling estimator yields similar improvements over the observed value $M^*=|E^*|$ as seen in estimating N.

Nevertheless, our preliminary work suggests that estimation of the number of edges M along the lines of the "leave-one-out" estimator is already a more challenging problem. At

some level (in a manner that can be made precise), the estimation of N is implicitly a problem closer to that of estimating the proportion of "species" unobserved (i.e., note the role of w^* in the L1O estimator), while the estimation of M is more explicitly a problem of estimating the number of "species" unobserved. The former type of problem is known to be easier than the latter type. Estimation of degrees k_i can be expected to be of even greater difficulty, given the comparatively low effective sample size per node i. Strategies that borrow strength across nodes likely will be necessary, such as the theoretical proposals in Ref. [29].

ACKNOWLEDGMENTS

F.V. was funded in part by the ACI Systémes et Sécurité, French Ministry of Research, as part of the MetroSec project. A.B. and L.D. are partially supported by the EU within the 6th Framework Programme under Contract No. 001907 (DELIS). Part of this work was performed while E.K. was with the LIAFA group at l'Université de Paris-7, with support from the CNRS. This work was supported in part by NSF Grant Nos. CCR-0325701 and DMS-0405202 and ONR Grant No. N000140310043 .

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[4] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2004).
[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[6] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[7] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994).
[8] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie, *Proceedings of IEEE INFOCOM*, San Francisco (2003).
[9] A. Clauset and C. Moore, Phys. Rev. Lett. **94**, 018701 (2005).
[10] T. Petermann and P. De Los Rios, Eur. Phys. J. B **38**, 201 (2004).
[11] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani, Phys. Rev. E **71**, 036135 (2005).
[12] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani, Theor. Comput. Sci. **355**, 6 (2006).
[13] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, *Proceedings of ACM STOC*, Baltimore (2003).
[14] J.-L. Guillaume and M. Latapy, *Proceedings of the IEEE INFOCOM*, Miami (2005).
[15] D. Alderson, J. C. Doyle, L. Li, and W. Willinger, Internet Math. **2**, 431 (2005).
[16] S. H. Lee, P.-J. Kim, and H. Jeong, Phys. Rev. E **73**, 016102 (2006).
[17] The National Laboratory for Applied Network Research (NLANR), sponsored by the National Science Foundation (see http://moat.nlanr.net/).
[18] The Cooperative Association for Internet Data Analysis (CAIDA), located at the San Diego Supercomputer Center (see http://www.caida.org/home/).
[19] Topology project, Electric Engineering and Computer Science Department, University of Michigan (http://topology.eecs.umich.edu/).
[20] SCAN project at the Information Sciences Institute (http://www.isi.edu/div7/scan/).
[21] Internet mapping project at Lucent Bell Labs (http://www.cs.bell-labs.com/who/ches/map/).
[22] Distributed Internet Measurements and Simulations (http://www.netdimes.org).
[23] S. K. Thompson, *Sampling* (Wiley, New York, 1992).
[24] O. Franks, Int. Statist. Rev. **48**, 33 (1980).
[25] H. Burch and B. Cheswick, IEEE Computer **32**, 97 (1999).
[26] J. Leguay, M. Latapy, T. Friedman, and K. Salamatian, *Proceedings of the 4th IFIP International Conference on Networking*, Waterloo, Canada (2005).
[27] B. Augustin, X. Cuvellier, T. Friedman, M. Latapy, C. Magnien, B. Orgogozo, F. Viger, and R. Teixera, *Proceedings of the 6th Internet Measurement Conference (IMC)*, Rio de Janiero, Brazil (2006).
[28] J. Bunge and M. Fitzpatrick, J. Am. Stat. Assoc. **88**, 364 (1993).
[29] C.-H. Zhang, Ann. Stat. **33**, 2022 (2005).
[30] The definition of betweenness centrality can straightforwardly be extended to routing models which do not consider shortest paths, with similar properties and conclusions. See, for example, Ref. [31] for a betweenness centrality measure based on random walks.
[31] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[32] K.-I. Goh, B. Kahng, and D. Kim, Physica A **318**, 72 (2003).
[33] G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley, New York, 2000).
[34] M. Barthélemy, Eur. Phys. J. B **38**, 163 (2004).
[35] K.-I. Goh, H. Jeong, B. Kahng, and D. Kim, Proc. Natl. Acad. Sci. U.S.A. **99**, 12 583 (2002).
[36] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*, *Society of Industrial and Applied Mathematics CBMS-NSF Monographs* (SIAM, Philadelphia, 1982), Vol. 38.
[37] F. Viger, A. Barrat, L. Dall'Asta, C.-H. Zhang, and E. D. Kolaczyk, eprint arXiv:cs.NI/0510007 (unpublished).
[38] P. Barford, A. Bestavros, J. Byers, and M. Crovella, *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, San Francisco (2001).
[39] I. J. Good, Biometrika **40**, 237 (1953).
[40] The Skitter project of CAIDA (see http://www.caida.org/tools/measurement/skitter/).
[41] The actual current number of destinations in the Skitter project is twice as big, but we restricted ourselves to the destinations that were reached in the process, for the sake of our models.
[42] R. Govindan and H. Tangmunarunkit, *Proceedings of the IEEE INFOCOM*, Tel-Aviv (2000).

[43] We again emphasize that the derivation of the estimators $\hat{N}_{\mathrm{RS}}$, and $\hat{N}_{L1O}$, does not make any explicit assumptions on the nature of the routing in the underlying network. Our use of shortest path routing here is only for the purpose of simulating in a simple way traceroutelike sampling. Numerical simulations using slightly more refined routing models [26] yield similar results.