Population size estimation and Internet link structure

Stephen E. Fienberg
Department of Statistics, Machine Learning Department, Cylab, and i-Lab
Carnegie Mellon University
Pittsburgh, PA 15213-3890
fienberg@stat.cmu.edu

Abraham D. Flaxman
Institute for Health Metrics and Evaluation, Department of Global Health
University of Washington
Seattle, WA
abie@uw.edu

Abstract

Traceroute sampling is a common approach for exploring the autonomous system (AS) graph of the Internet. It provides samples of links between autonomous systems, but these links are not drawn uniformly at random from all possible links. Rather, the rules that each AS uses are idiosyncratic and emergent. Here, we are interested in using the data from traceroute sampling to estimate the degree distribution of the network, a quantity of common interest in network modeling more broadly. We link these ideas to the methodology of multiple-recapture estimation of the size of a closed population using log-linear models. We apply our approach to produce new estimates of the degree distribution of the AS graph, and to provide further evidence that the degree distribution does indeed have heavy tails.

1 Introduction

The Internet is a decentralized tangle of routers, with a natural clustering where each set of routers under the common management are grouped as an Autonomous System (AS). The links between the routers are not recorded in any central repository, and so obtaining accurate data on the link structure of the AS graph is an important step in understanding the Internet. There are three techniques for finding large sets of edges in the AS graph: the WHOIS database, BGP tables, and traceroute sampling. No approach is clearly superior, and the results of the different approaches are compared in detail in [16].

The present paper focuses on traceroute sampling, which consists of recording the paths that packets follow when they are sent from monitor nodes to target nodes, and merging all of these paths to produce an approximation of the AS graph. A seminal analysis using both traceroute sampling and BGP tables concluded that the AS graph degree distribution follows a power-law (meaning that the number of AS of degree k is proportional to $k^{-\alpha}$ for a wide range of k values) [7]. This caused a shift in simulation methodology for evaluating network algorithms and also contributed to the avalanche of research developing new network models that produce power-law degree distributions.

Lakhina et al. [14, 19] called into question the true nature of the AS graph degree distribution by computer experiments on synthetic graphs showing that if the sets of monitor and target nodes are too small, then a simple model of traceroute sampling produces a power-law degree distribution, even when the underlying graph has a tightly concentrated degree distribution. Mathematically rigorous follow-up work proved that in many non-power-law graphs, including random regular graphs, an

idealized model of traceroute sampling yields power-law degree distributions [5, 1]. See also the more general result in [21] and the discussion in [12].

Flaxman and Vera [10] showed that an adhoc modification of the Petersen estimate from multiple-recapture population estimation provides an unbiased estimator for the degree distribution of Erdős-Rényi graphs under a simple model of traceroute sampling for the AS graph. The present paper revisits their approach and draws more heavily from the theory of multiple-recapture population estimation. We use log-linear modeling techniques to estimate the AS graph degree distribution directly.

Viger, Barrat, Dall'Asta, Zhang, and Kolaczyk [22] applied a related technique from statistics known as the frequencies of frequencies, or species problem, to reduce the bias of traceroute sampling. The problem of correcting bias in sampled networks has a long history in sociological applications of network modeling, although the biases in that domain seem somewhat different. See the surveys by Frank, Klovdahl, or Salganik and Heckathorn for an overview [11, 13, 20].

2 Traceroute Sampling and a Shortest Path Sampling

We formalize traceroute sampling as follows. Let $G = (V, E)$ be the AS graph, where V is the set of autonomous systems, and undirected edges in E correspond to one hop connectivity between AS. Traceroute sampling from a set of k monitors consists of sending packets to m target nodes from each monitor and recording the paths that the packets follow en route. The union of the paths from monitor i yields a sampled graph G_i for $i = 1, \ldots, k$, and each G_i contains at least m nodes and edges.

Conceptually, we can think of the paths discovered in traceroute sampling as approximations to the shortest paths between the monitor and the target in the underlying AS graph G, akin to the paths for packets in Milgram's original small-world experiments. (See [13, 12]). There is now general agreement that the paths that data actually take are not the shortest paths (see [15] for a discussion of this approximation). Approximating the actual paths of traceroute sampling with the shortest path in the graph is convenient for simulations, however. We use this approach in Section 4.

3 Estimation Technique

In contrast to the indirect approach taken in Flaxman and Vera [10], we estimate the degree distribution of the traceroute-sampled graph directly by estimating the number of nodes with degree at least d for a range of values of d. This arises naturally in the context of the log-linear model approach to multiple-recapture estimation of a population total.

In order to apply multiple-recapture techniques to the traceroute-sampled AS graph, we must come up with recapture phases. We do this by identifying the edges discovered by traceroute samples from (potentially overlapping) sets of monitors with recapture phases. This generalizes the approach of Flaxman and Vera [10], who considered different monitors as different phases. Using sets of monitors as phases allows a balance between the computational cost of using many recapture phases and the potential effects on accuracy from taking the union of edge sets discovered by different monitors.

Let G be a graph, and let $i_1, i_2, \ldots, i_k$ denote the k monitor nodes in G. Let G_S be the subgraph consisting of the union of all routes discovered when sending packets from any monitor $i \in S$ to all m nodes in the target set. Flaxman and Vera considered disjoint singleton sets $S_j = \{i_j\}$ and examined the degree of each observed node in each subgraph and in the pairwise intersections of them. They used a combination of the Petersen estimators for a simple capture-recapture estimate for each pair in order to estimate the population degree distribution by looking at the estimates of the degree node by node.

We instead considered the complementary cumulative distribution of the degree distribution directly. We work with the 2^J contingency tables formed by fixing J (not necessarily disjoint) sets of monitor nodes $(S_1, S_2, \ldots, S_J)$ and enumerating the nodes in the G_{S_j}'s with degree at least d, for $d = 2, 3, \ldots, n$. In each such table, we do not get to observe the count of nodes of degree at least d,

which are in none of the J subgraphs. We use multiple-recapture estimation to estimate the missing cell in each of these tables.

One of the attractive features of the multiple-recapture technology is that all of the maximum likelihood estimates of the number of nodes of degree at least d, N_d, take the form:

$$\widehat{N_d} = n_d + \text{estimate of missing cell},$$

where n_d is the observed number of nodes of degree at least d that appear in at least one of the J subgraphs. We considered including one subgraph with all monitors, i.e., G_S with $S = \{i_1, \ldots, i_k\}$, which results in n_d matching the empirical cdf from the union of the traceroute samples. However, this did not perform as well as the disjoint clusters used in the simulation experiments below.

If we estimate the missing cell in each contingency table separately, nothing prohibits the estimates $\widehat{N_d}$ from increasing as d increases. But then when we difference the $\widehat{N_d}$ to get the degree distribution we could in principle get negative probability estimates. To prevent this inconsistency, we consider a decreasing-constrained variant of the log-linear model that fits models to all contingency tables simultaneously and adds a constraint that $\widehat{N_d}$ is decreasing as a function of d.

We have focused for empirical tests on simulated networks, with number of nodes, edges, monitors, and targets chosen similarly to that found in the traceroute-sampled AS graph collected in March 2003 by the skitter project. We considered two naive approaches—estimates from the union of all edges discovered, and Petersen estimates derived from two disjoint-monitor-set subgraphs G_{S_1} and G_{S_2}. We also considered log-linear models with first-order and second-order interactions, described in more detail below, and an extension of the second-order log-linear model, which adds a constraint that the degree distribution decreases, as described above.

The key feature of the log-linear model approach is to focus on dependencies in the traceroutes. There is no reason to believe that the events "monitor i observes link j" are independent. Indeed, when shortest-path routing is used (as an approximation of BGP routing), these events are highly dependent.

Because of these dependencies, we used the log-linear model approach, which preserves marginal sums from the 2^J tables. The one-way margins directly count the number of nodes with degree at least d in each subgraph. If we preserve only one-way margins, then when we fit only these marginal sums, we are in effect assuming independence of degree-at-least d nodes appearing in each subgraph, $\{G_{S_j}\}$. The two-way marginal totals capture the first-order dependencies among the subgraphs, providing a way to correct the erroneous assumption enforced by the first-order model. For a detailed description of the estimation methodology, see Fienberg [9] and Bishop et al. [3], and for extensions to this approach that add in features of heterogeneity among the nodes, see [8, 17].

In short, the first-order log-linear model of a 2^J contingency table introduces parameters μ_j for each of the J lists. It encodes the appearance/non-appearance of a node in list j as $s_j = \pm 1$, and then models the probability of the node appearing in a particular cell of the contingency table by $\log p_s = \sum_{j=1}^{J} s_j \mu_j$. The cell counts are then modeled by a multinomial distribution, with probabilities given by the p_s's.

The second-order log-linear model (with all two-factor interactions) augments the first-order model with additional parameters $\mu_{i,j}$ for each pair of lists. These parameters are included in the probability of a node appearing in a cell as a signed sum, $\log p'_s = \log p_s + \sum_{i \neq j} s_i s_j \mu_{i,j}$.

To realize the constraint that the $\widehat{N_d}$ values decrease as a function of d, in the second-order decreasing log-linear model, we added a quadratic penalty function on the difference $D_d = (\widehat{N}_{d-1} - \widehat{N}_d)^+$.

We fitted these models using Python/PyMC, by generating initial values with likelihood maximization and drawing 5,000 samples with Adaptive Metropolis MCMC (after discarding 5,000 burn-in samples) to obtain posterior means and uncertainty intervals on $\widehat{N_d}$.

4 Validation with Simulated Data

This section describes the results of a series of computer experiments conducted to investigate how well $\widehat{N_d}$ approximates the true degree distribution.

We considered three different distributions for random graphs—the Erdős-Rényi model, the Preferential Attachment model, and the edge-percolated random geometric graph.

For each graph, we set edge e to be of length $1+\eta_e$, where η_e is selected uniformly from the interval $[-1/n, 1/n]$, where n is the number of vertices. This ensures that there are not multiple shortest paths between pairs of vertices. We approximated the path that data takes from a monitor to a target node by the shortest path. This follows the experimental design used in [14].

For each graph distribution, we estimated the number of vertices with degree at least d using first-order and second-order log-linear models, as well as the decreasing-constrained second-order log-linear model. We also calculated the Petersen estimate on two subgraphs and the naive estimator formed by considering the union of the edges discovered. The naive estimator, as the discussion above makes clear, provides a form of lower bound on sensible estimates.

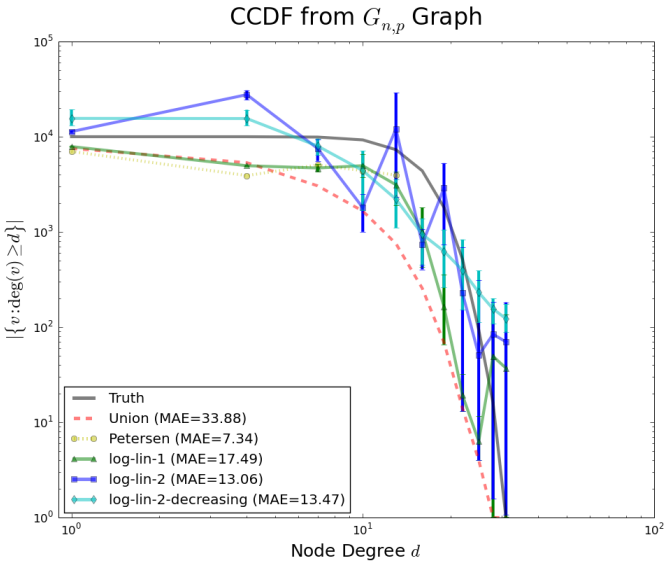
To measure the quality of each estimate, we calculated the relative median absolute error (MAE) of the log of $\widehat{N_d}$ for each estimate compared to the log of the true degree distribution,

$$\text{err}_i = \text{Median}\left(\left|100\frac{\log \widehat{N_d} - \log N_d}{\log N_d}\right|\right)$$

4.1 Random Graph, $G_{n,m}$

The Erdős-Rényi distribution of graphs, $G_{n,m}$, involves choosing a graph uniformly at random from all graphs with n vertices and m edges [6]. It was not developed to model real-world graphs, but it is analytically tractable and can provide insight into the behavior of more realistic graph models. We generated instances with $n = 10,000$ and $m = 7.5n$ (to yield average degree 15), and simulated traceroute sampling using shortest paths between 24 monitor nodes chosen at random without replacement and $1,000$ target nodes also chosen at random without replacement. We used 4 sets of 6 monitors each for the subgraphs. The uncertainty interval comes from running 64 independent replicates of the experiment.

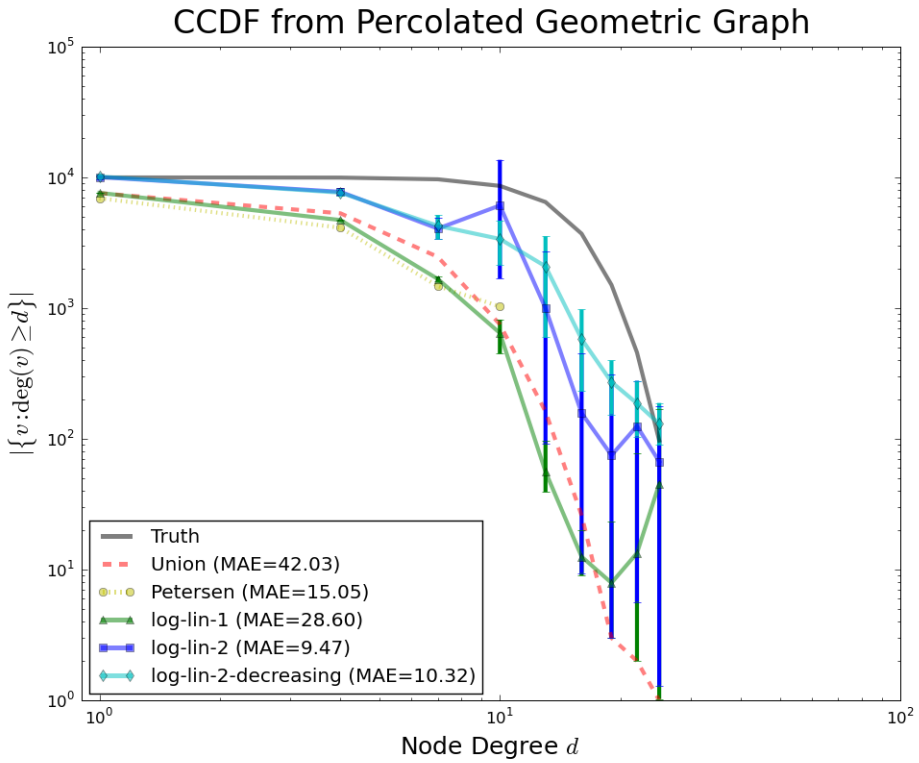
Method	Relative MAE %	(95% UI)
Union	27.80	(25, 35)
Petersen	8.07	(4, 12)
Log-Linear 1	11.53	(8, 17)
Log-Linear 2	8.96	(3, 16)
Log-Linear 2 Decreasing	7.75	(2, 14)



4.2 Edge-Percolated Random Geometric Graph, $G(\mathcal{X}; r)$

To investigate the performance of the bias-reduction technique on graphs with clustering, we examined random geometric graphs $G(\mathcal{X}; r)$. These graphs are formed by selecting a set of n points independently and uniformly at random from the unit square, and linking two points with an edge if and only if they are within ℓ_2 distance r (for a detailed treatment, see [18]). The edge percolation then selected edges from this random geometric graph independently at random with probability p. We generated instances with $n = 10,000$, $p = .1$, and r chosen to yield average degree 15, and simulated traceroute sampling using shortest paths between 24 monitor nodes chosen at random without replacement and $1,000$ target nodes also chosen at random without replacement. We used 4 sets of 6 monitors each for the subgraphs. The uncertainty interval comes from running 64 independent replicates of the experiment.

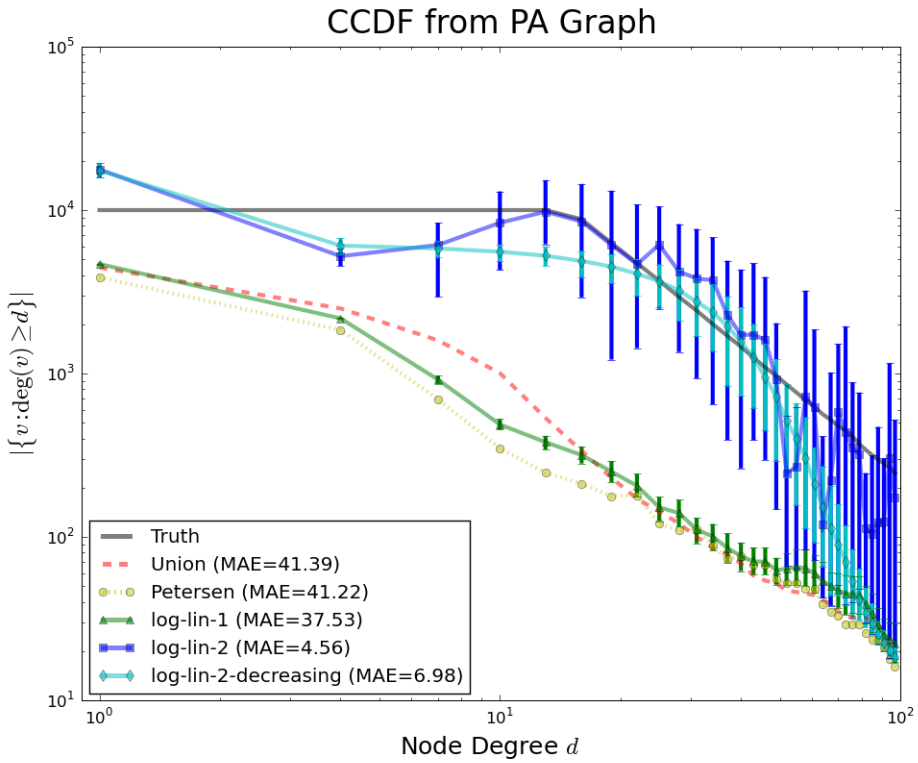
Method	Relative MAE %	(95% UI)
Union	32.92	(27, 51)
Petersen	17.01	(14, 24)
Log-Linear 1	31.91	(26, 41)
Log-Linear 2	13.29	(8, 23)
Log-Linear 2 Decreasing	13.38	(10, 19)



4.3 Preferential Attachment Graph

The preferential attachment (PA) graph was proposed for a model of the Internet and the World Wide Web by Barabási and Albert in [2], and this has generated a large body of subsequent research, although the validity of the model as a representation of the router graph or the AS graph has been questioned (see, for example, [4]). We generated instances with $n = 10,000$ and minimum degree 15, and simulated traceroute sampling using shortest paths between 24 monitor nodes chosen at random without replacement and $1,000$ target nodes also chosen at random without replacement. We used 4 sets of 6 monitors each for the subgraphs. The uncertainty interval comes from running 64 independent replicates of the experiment.

Method	Relative MAE %	(95% UI)
Union	42.87	(40, 46)
Petersen	41.93	(36, 50)
Log-Linear 1	39.92	(37, 44)
Log-Linear 2	7.64	(5, 15)
Log-Linear 2 Decreasing	14.85	(6, 26)

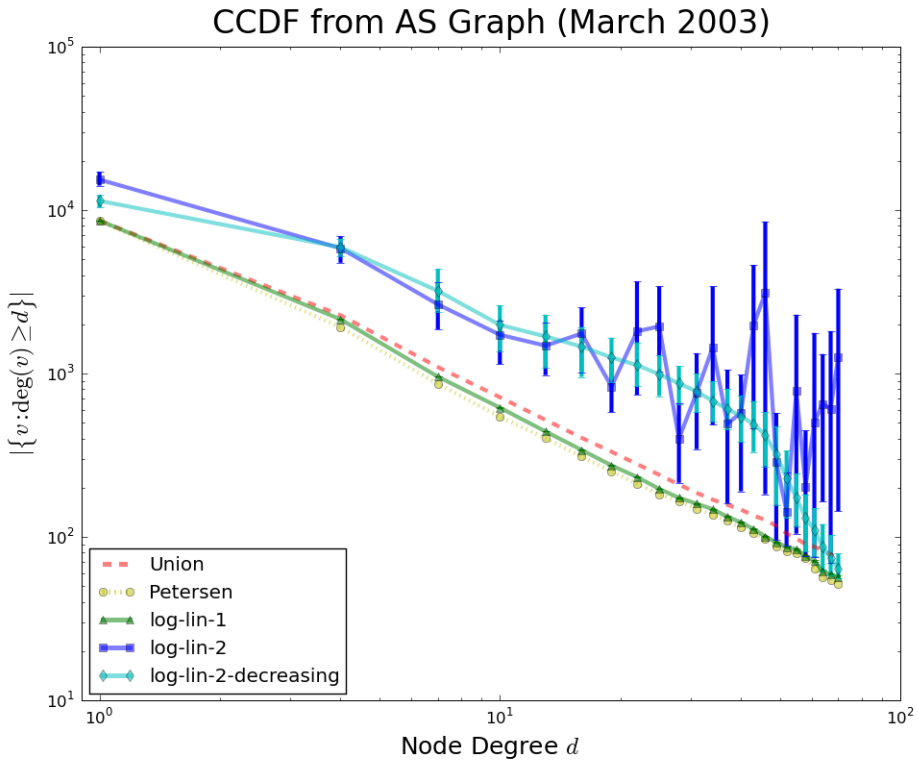


5 Recapture estimates for the AS graph

This section reports on the results of applying the bias-reduction technique to traceroute-sampled data from the CAIDA skitter project.

Mahadevan et al. [16] provide a detailed analysis of CAIDA skitter data from March 2004. We followed their methodology, and, in particular, we aggregated the routes observed over the course of a month (from daily graphs provided by CAIDA). We also removed all AS sets, multi-origin AS, and private AS, and discarded all indirect links.

The following graph shows results of applying the bias-reduction technique to the March 2003 skitter data. There are 24 AS in the monitor set. Each monitor sees around 7,000 nodes and around 12,000 edges. Since there are 24 AS in the monitor set, the contingency table for log-linear estimation with singleton clusters has 2^{24} cells. To decrease computation time, we consider 4 sets of 6 monitors each.



6 Conclusion

In this paper, we have built on the notion from Flaxman and Vera on the link between bias correction of the degree distribution for the AS graph and the use of a variant of the traditional Petersen estimate from simple capture-recapture settings. But unlike those authors, we have used the methodology from multiple-recapture settings involving log-linear models applied to a sequence of constructed contingency tables in order to get direct estimates of the degree distribution. By building in dependencies among the collections of traceroute subgraphs, we appear to have improved estimates (relative to prior methods) of the degree distribution for three types of standard test graphs: (1) the Erdős-Rényi Random Graph, $G_{n,m}$, (2) the Preferential Attachment Graph, and (3) the Edge-Percolated Random Geometric Graph, $G(\mathcal{X}; r, p)$.

We view these empirical results as preliminary since we have done only a limited set of experiments in each setting. Moreover, the simple second-order log-linear interaction model would appear to capture information on a restricted set of features from the simulated graphs. We expect the use of somewhat more elaborate log-linear styled models will do a better job of capturing aspects of heterogeneity and higher order dependence among subgraphs generated by traceroute paths and that hierarchically smoothed versions of these models will perform even better.

Our use of multiple-recapture methodology is heuristically appealing because of the simple interpretation of the minimal sufficient statistics of the first- and second-order log-linear models (see similar heuristics in [22]). We need more formal probabilistic links between these and other multiple-recapture models and common classes of network models to provide our work with a firmer theoretical foundation.

References

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *STOC '05: Proceedings of the thirty-*

seventh annual ACM symposium on Theory of computing, pages 694–703, New York, NY, USA, 2005. ACM Press.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.

[4] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. The origin of power laws in Internet topologies revisited. In *INFOCOM 2002, Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, Proceedings*, volume 2, pages 608–617, 2002.

[5] A. Clauset and C. Moore. Accuracy and scaling phenomena in internet mapping. *Physical Review Letters*, 94(1):018701, 2005.

[6] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA, 1999. ACM Press.

[8] S. Fienberg, M. Johnson, and B. Junker. Classical multilevel and bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162:383–405(23), 1999.

[9] S. E. Fienberg. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59(3):591–603, 1972.

[10] A. D. Flaxman and J. Vera. Bias reduction in traceroute sampling: towards a more accurate map of the Internet. In *Proceedings of the 5th International Workshop on Algorithms and Models for the Web-Graph (WAW)*, pages 1–15, 2007. To appear in Web Algorithms Workshop 2007.

[11] O. Frank. A survey of statistical methods for graph analysis. *Sociological Methodology*, 12:110–155, 1981.

[12] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[13] A. S. Klovdahl. *The Small World (in honor of Stanley Milgram)*, chapter Urban social networks: Some methodological problems and possibilities. ABLEX, 1989.

[14] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in ip topology measurements. In *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, volume 1, pages 332–341. IEEE, April 2003.

[15] J. Leguay, M. Latapy, T. Friedman, and K. Salamatian. Describing and simulating internet routes. In *4th International IFIP-TC6 Networking Conference, Waterloo, Canada, May 2-6, 2005. Proceedings*, volume 3462 of *Lecture Notes in Computer Science*, pages 659–670, 2005.

[16] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, k c claffy, and A. Vahdat. The internet AS-level topology: three data sources and one definitive metric. *SIGCOMM Comput. Commun. Rev.*, 36(1):17–26, 2006.

[17] D. Manrique-Vallier and S. E. Fienberg. Population size estimation using individual level mixture models. *Biometrical Journal*, 50(6):1051–1063, 2008.

[18] M. Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

[19] T. Petermann and P. D. L. Rios. Exploration of scale-free networks. *European Physical Journal B*, 38:201–204, 2004.

[20] M. J. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-drive sampling. *Sociological Methodology*, 34:193–239, 2004.

[21] M. Stumpf, C.Wiuf, and R. May. Subnets of scale-free networks are not scale-free. *Proceedings of the National Academy of Sciences*, 103:7566–7570, 2004.

[22] F. Viger, A. Barrat, L. Dall'Asta, C. Zhang, and E. Kolaczyk. Network Inference from TraceRoute Measurements: Internet Topology 'Species'. *Phys. Rev. E*, 75(056111), 2007.