Anthony Bonato
Fan R.K. Chung (Eds.)

LNCS 4863

Algorithms and Models for the Web-Graph

5th International Workshop, WAW 2007
San Diego, CA, USA, December 2007
Proceedings

Springer

Anthony Bonato Fan R.K. Chung (Eds.)

Algorithms and Models for the Web-Graph

5th International Workshop, WAW 2007
San Diego, CA, USA, December 11-12, 2007
Proceedings

Volume Editors

Anthony Bonato
Wilfrid Laurier University, Department of Mathematics
Waterloo, ON, N2L 3C5, Canada
E-mail: abonato@rogers.com

Fan R.K. Chung
University of California, San Diego, Department of Mathematics
La Jolla, CA 92093-0112, USA
E-mail: fan@math.ucsd.edu

Library of Congress Control Number: 2007939821

CR Subject Classification (1998): F.2, G.2, H.4, H.3, C.2, H.2.8, E.1

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-77003-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-77003-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12197316 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4863

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Preface

This volume constitutes the refereed proceedings of the Fifth Workshop on Algorithms and Models for the Web-Graph, WAW 2007, held in San Diego in December 2007. The proceedings consist of 18 revised papers (13 regular papers and 5 short papers) which were reviewed and selected from a large pool of submissions. The papers address a wide variety of topics related to the study of the Web-graph such as random graph models for the Web-graph, PageRank analysis and computation, decentralized search, local partitioning algorithms, and traceroute sampling.

The Web-graph has been the focal point of a tremendous amount of research for more than a decade. The view of the Web as a graph has great practical importance and has also generated much interesting theoretical work. A goal of the 2007 Workshop was to present state-of-the art research on both the applications and theory of the Web-graph. Our hope is that the papers presented here will help stimulate new and exciting avenues of research on the Web-graph.

December 2007

Anthony Bonato
Fan Chung Graham

Organization

Executive Committee

Conference Chair	Ronald Graham (University of California, San Diego, USA)
Local Arrangements Chair	Tara Javidi (University of California, San Diego, USA)
Program Committee Co-chair	Anthony Bonato (Wilfrid Laurier University, Canada)
Program Committee Co-chair	Fan Chung Graham (University of California, San Diego, USA)
Program Committee Co-chair	Tara Javidi (University of California, San Diego, USA)

Organizing Committee

Andrei Broder, (Yahoo! Research, USA)
Fan Chung Graham (University of California, San Diego, USA)
Jeannette Janssen, (Dalhousie University, Canada)
Tara Javidi (University of California, San Diego, USA)
Lincoln Lu (University of South Carolina, USA)

Program Committee

Dimitris Achlioptas, (University of California, Santa Cruz, USA)
Colin Cooper, (King's College London, UK)
Anthony Bonato (Wilfrid Laurier University, Canada)
Alan Frieze (Carnegie Mellon University, USA)
Michael Goodrich, (University of California, Irvine, USA)
Fan Chung Graham (University of California, San Diego, USA)
Jeannette Janssen, (Dalhousie University, Canada)
Tara Javidi (University of California, San Diego, USA)
Ravi Kumar (Yahoo! Research, USA)
Kevin Lang, (Yahoo! Research, USA)
Stefano Leonardi (Università di Roma, Italy)
Lincoln Lu (University of South Carolina, USA)
Milena Mihail (Georgia Institute of Technology, USA)
Michael Mitzenmacher (Harvard University, USA)
Muthu Muthukrishnan (Rutgers University and Google Inc., USA)
Joel Spencer (New York University, USA)
Walter Willinger (AT&T Research, USA)

Sponsoring Institutions

California Institute for Telecommunications and Information Technology
Google Inc.
Yahoo! Research
National Science Foundation
Springer *Lecture Notes in Computer Science*
University of California, San Diego

Lecture Notes in Computer Science

Sublibrary 1: Theoretical Computer Science and General Issues

For information about Vols. 1– 4497
please contact your bookseller or Springer

Vol. 4863: A. Bonato, F.R.K. Chung (Eds.), Algorithms and Models for the Web-Graph. X, 217 pages. 2007.

Vol. 4847: M. Xu, Y. Zhan, J. Cao, Y. Liu (Eds.), Advanced Parallel Processing Technologies. XIX, 767 pages. 2007.

Vol. 4846: I. Cervesato (Ed.), Advances in Computer Science - ASIAN 2007. XI, 313 pages. 2007.

Vol. 4838: T. Masuzawa, S. Tixeuil (Eds.), Stabilization, Safety, and Security of Distributed Systems. XIII, 409 pages. 2007.

Vol. 4783: J. Holub, J. Žďárek (Eds.), Implementation and Application of Automata. XIII, 324 pages. 2007.

Vol. 4782: R. Perrott, B.M. Chapman, J. Subhlok, R.F. de Mello, L.T. Yang (Eds.), High Performance Computing and Communications. XIX, 823 pages. 2007.

Vol. 4771: T. Bartz-Beielstein, M.J. Blesa Aguilera, C. Blum, B. Naujoks, A. Roli, G. Rudolph, M. Sampels (Eds.), Hybrid Metaheuristics. X, 202 pages. 2007.

Vol. 4770: V.G. Ganzha, E.W. Mayr, E.V. Vorozhtsov (Eds.), Computer Algebra in Scientific Computing. XIII, 460 pages. 2007.

Vol. 4763: J.-F. Raskin, P.S. Thiagarajan (Eds.), Formal Modeling and Analysis of Timed Systems. X, 369 pages. 2007.

Vol. 4746: A. Bondavalli, F. Brasileiro, S. Rajsbaum (Eds.), Dependable Computing. XV, 239 pages. 2007.

Vol. 4743: P. Thulasiraman, X. He, T.L. Xu, M.K. Denko, R.K. Thulasiram, L.T. Yang (Eds.), Frontiers of High Performance Computing and Networking ISPA 2007 Workshops. XXIX, 536 pages. 2007.

Vol. 4742: I. Stojmenovic, R.K. Thulasiram, L.T. Yang, W. Jia, M. Guo, R.F. de Mello (Eds.), Parallel and Distributed Processing and Applications. XX, 995 pages. 2007.

Vol. 4739: R. Moreno Díaz, F. Pichler, A. Quesada Arencibia (Eds.), Computer Aided Systems Theory – EUROCAST 2007. XIX, 1233 pages. 2007.

Vol. 4736: S. Winter, M. Duckham, L. Kulik, B. Kuipers (Eds.), Spatial Information Theory. XV, 455 pages. 2007.

Vol. 4732: K. Schneider, J. Brandt (Eds.), Theorem Proving in Higher Order Logics. IX, 401 pages. 2007.

Vol. 4731: A. Pelc (Ed.), Distributed Computing. XVI, 510 pages. 2007.

Vol. 4726: N. Ziviani, R. Baeza-Yates (Eds.), String Processing and Information Retrieval. XII, 311 pages. 2007.

Vol. 4711: C.B. Jones, Z. Liu, J. Woodcock (Eds.), Theoretical Aspects of Computing – ICTAC 2007. XI, 483 pages. 2007.

Vol. 4710: C.W. George, Z. Liu, J. Woodcock (Eds.), Domain Modeling and the Duration Calculus. XI, 237 pages. 2007.

Vol. 4708: L. Kučera, A. Kučera (Eds.), Mathematical Foundations of Computer Science 2007. XVIII, 764 pages. 2007.

Vol. 4707: O. Gervasi, M.L. Gavrilova (Eds.), Computational Science and Its Applications – ICCSA 2007, Part III. XXIV, 1205 pages. 2007.

Vol. 4706: O. Gervasi, M.L. Gavrilova (Eds.), Computational Science and Its Applications – ICCSA 2007, Part II. XXIII, 1129 pages. 2007.

Vol. 4705: O. Gervasi, M.L. Gavrilova (Eds.), Computational Science and Its Applications – ICCSA 2007, Part I. XLIV, 1169 pages. 2007.

Vol. 4703: L. Caires, V.T. Vasconcelos (Eds.), CONCUR 2007 – Concurrency Theory. XIII, 507 pages. 2007.

Vol. 4700: C.B. Jones, Z. Liu, J. Woodcock (Eds.), Formal Methods and Hybrid Real-Time Systems. XVI, 539 pages. 2007.

Vol. 4699: B. Kågström, E. Elmroth, J. Dongarra, J. Waśniewski (Eds.), Applied Parallel Computing. XXIX, 1192 pages. 2007.

Vol. 4698: L. Arge, M. Hoffmann, E. Welzl (Eds.), Algorithms – ESA 2007. XV, 769 pages. 2007.

Vol. 4697: L. Choi, Y. Paek, S. Cho (Eds.), Advances in Computer Systems Architecture. XIII, 400 pages. 2007.

Vol. 4688: K. Li, M. Fei, G.W. Irwin, S. Ma (Eds.), Bio-Inspired Computational Intelligence and Applications. XIX, 805 pages. 2007.

Vol. 4684: L. Kang, Y. Liu, S. Zeng (Eds.), Evolvable Systems: From Biology to Hardware. XIV, 446 pages. 2007.

Vol. 4683: L. Kang, Y. Liu, S. Zeng (Eds.), Advances in Computation and Intelligence. XVII, 663 pages. 2007.

Vol. 4681: D.-S. Huang, L. Heutte, M. Loog (Eds.), Advanced Intelligent Computing Theories and Applications. XXVI, 1379 pages. 2007.

Vol. 4672: K. Li, C. Jesshope, H. Jin, J.-L. Gaudiot (Eds.), Network and Parallel Computing. XVIII, 558 pages. 2007.

Vol. 4671: V.E. Malyshkin (Ed.), Parallel Computing Technologies. XIV, 635 pages. 2007.

Vol. 4669: J.M. de Sá, L.A. Alexandre, W. Duch, D. Mandic (Eds.), Artificial Neural Networks – ICANN 2007, Part II. XXXI, 990 pages. 2007.

Vol. 4668: J.M. de Sá, L.A. Alexandre, W. Duch, D. Mandic (Eds.), Artificial Neural Networks – ICANN 2007, Part I. XXXI, 978 pages. 2007.

Vol. 4666: M.E. Davies, C.J. James, S.A. Abdallah, M.D. Plumbley (Eds.), Independent Component Analysis and Blind Signal Separation. XIX, 847 pages. 2007.

Vol. 4665: J. Hromkovič, R. Královič, M. Nunkesser, P. Widmayer (Eds.), Stochastic Algorithms: Foundations and Applications. X, 167 pages. 2007.

Vol. 4664: J. Durand-Lose, M. Margenstern (Eds.), Machines, Computations, and Universality. X, 325 pages. 2007.

Vol. 4661: U. Montanari, D. Sannella, R. Bruni (Eds.), Trustworthy Global Computing. X, 339 pages. 2007.

Vol. 4649: V. Diekert, M.V. Volkov, A. Voronkov (Eds.), Computer Science – Theory and Applications. XIII, 420 pages. 2007.

Vol. 4647: R. Martin, M.A. Sabin, J.R. Winkler (Eds.), Mathematics of Surfaces XII. IX, 509 pages. 2007.

Vol. 4646: J. Duparc, T.A. Henzinger (Eds.), Computer Science Logic. XIV, 600 pages. 2007.

Vol. 4644: N. Azémard, L. Svensson (Eds.), Integrated Circuit and System Design. XIV, 583 pages. 2007.

Vol. 4641: A.-M. Kermarrec, L. Bougé, T. Priol (Eds.), Euro-Par 2007 Parallel Processing. XXVII, 974 pages. 2007.

Vol. 4639: E. Csuhaj-Varjú, Z. Ésik (Eds.), Fundamentals of Computation Theory. XIV, 508 pages. 2007.

Vol. 4638: T. Stützle, M. Birattari, H. H. Hoos (Eds.), Engineering Stochastic Local Search Algorithms. X, 223 pages. 2007.

Vol. 4630: H.J. van den Herik, P. Ciancarini, H.H.L.M.(J.) Donkers (Eds.), Computers and Games. XII, 283 pages. 2007.

Vol. 4628: L.N. de Castro, F.J. Von Zuben, H. Knidel (Eds.), Artificial Immune Systems. XII, 438 pages. 2007.

Vol. 4627: M. Charikar, K. Jansen, O. Reingold, J.D.P. Rolim (Eds.), Approximation, Randomization, and Combinatorial Optimization. XII, 626 pages. 2007.

Vol. 4624: T. Mossakowski, U. Montanari, M. Haveraaen (Eds.), Algebra and Coalgebra in Computer Science. XI, 463 pages. 2007.

Vol. 4623: M. Collard (Ed.), Ontologies-Based Databases and Information Systems. X, 153 pages. 2007.

Vol. 4621: D. Wagner, R. Wattenhofer (Eds.), Algorithms for Sensor and Ad Hoc Networks. XIII, 415 pages. 2007.

Vol. 4619: F. Dehne, J.-R. Sack, N. Zeh (Eds.), Algorithms and Data Structures. XVI, 662 pages. 2007.

Vol. 4618: S.G. Akl, C.S. Calude, M.J. Dinneen, G. Rozenberg, H.T. Wareham (Eds.), Unconventional Computation. X, 243 pages. 2007.

Vol. 4616: A.W.M. Dress, Y. Xu, B. Zhu (Eds.), Combinatorial Optimization and Applications. XI, 390 pages. 2007.

Vol. 4614: B. Chen, M. Paterson, G. Zhang (Eds.), Combinatorics, Algorithms, Probabilistic and Experimental Methodologies. XII, 530 pages. 2007.

Vol. 4613: F.P. Preparata, Q. Fang (Eds.), Frontiers in Algorithmics. XI, 348 pages. 2007.

Vol. 4600: H. Comon-Lundh, C. Kirchner, H. Kirchner (Eds.), Rewriting, Computation and Proof. XVI, 273 pages. 2007.

Vol. 4599: S. Vassiliadis, M. Bereković, T.D. Hämäläinen (Eds.), Embedded Computer Systems: Architectures, Modeling, and Simulation. XVIII, 466 pages. 2007.

Vol. 4598: G. Lin (Ed.), Computing and Combinatorics. XII, 570 pages. 2007.

Vol. 4596: L. Arge, C. Cachin, T. Jurdziński, A. Tarlecki (Eds.), Automata, Languages and Programming. XVII, 953 pages. 2007.

Vol. 4595: D. Bošnački, S. Edelkamp (Eds.), Model Checking Software. X, 285 pages. 2007.

Vol. 4590: W. Damm, H. Hermanns (Eds.), Computer Aided Verification. XV, 562 pages. 2007.

Vol. 4588: T. Harju, J. Karhumäki, A. Lepistö (Eds.), Developments in Language Theory. XI, 423 pages. 2007.

Vol. 4583: S.R. Della Rocca (Ed.), Typed Lambda Calculi and Applications. X, 397 pages. 2007.

Vol. 4580: B. Ma, K. Zhang (Eds.), Combinatorial Pattern Matching. XII, 366 pages. 2007.

Vol. 4576: D. Leivant, R. de Queiroz (Eds.), Logic, Language, Information and Computation. X, 363 pages. 2007.

Vol. 4547: C. Carlet, B. Sunar (Eds.), Arithmetic of Finite Fields. XI, 355 pages. 2007.

Vol. 4546: J. Kleijn, A. Yakovlev (Eds.), Petri Nets and Other Models of Concurrency – ICATPN 2007. XI, 515 pages. 2007.

Vol. 4545: H. Anai, K. Horimoto, T. Kutsia (Eds.), Algebraic Biology. XIII, 379 pages. 2007.

Vol. 4533: F. Baader (Ed.), Term Rewriting and Applications. XII, 419 pages. 2007.

Vol. 4528: J. Mira, J.R. Álvarez (Eds.), Nature Inspired Problem-Solving Methods in Knowledge Engineering, Part II. XXII, 650 pages. 2007.

Vol. 4527: J. Mira, J.R. Álvarez (Eds.), Bio-inspired Modeling of Cognitive Tasks, Part I. XXII, 630 pages. 2007.

Vol. 4525: C. Demetrescu (Ed.), Experimental Algorithms. XIII, 448 pages. 2007.

Vol. 4514: S.N. Artemov, A. Nerode (Eds.), Logical Foundations of Computer Science. XI, 513 pages. 2007.

Vol. 4513: M. Fischetti, D.P. Williamson (Eds.), Integer Programming and Combinatorial Optimization. IX, 500 pages. 2007.

Vol. 4510: P. Van Hentenryck, L.A. Wolsey (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. X, 391 pages. 2007.

Vol. 4507: F. Sandoval, A.G. Prieto, J. Cabestany, M. Graña (Eds.), Computational and Ambient Intelligence. XXVI, 1167 pages. 2007.

Vol. 4502: T. Altenkirch, C. McBride (Eds.), Types for Proofs and Programs. VIII, 269 pages. 2007.

Vol. 4501: J. Marques-Silva, K.A. Sakallah (Eds.), Theory and Applications of Satisfiability Testing – SAT 2007. XI, 384 pages. 2007.

Table of Contents

Bias Reduction in Traceroute Sampling – Towards a More Accurate Map of the Internet

Abraham D. Flaxman[1] and Juan Vera[2]

[1] Microsoft Research
Redmond, WA
abie@microsoft.com
[2] Georgia Institute of Technology
Atlanta, GA
jvera@cc.gatech.edu

Abstract. Traceroute sampling is an important technique in exploring the internet router graph and the autonomous system graph. Although it is one of the primary techniques used in calculating statistics about the internet, it can introduce bias that corrupts these estimates. This paper reports on a theoretical and experimental investigation of a new technique to reduce the bias of traceroute sampling when estimating the degree distribution. We develop a new estimator for the degree of a node in a traceroute-sampled graph; validate the estimator theoretically in Erdős-Rényi graphs and, through computer experiments, for a wider range of graphs; and apply it to produce a new picture of the degree distribution of the autonomous system graph.

1 Introduction

The internet is quite a mysterious network. It is a huge and complex tangle of routers, wired together by millions of edges. To understand this *router graph* is quite a challenge, one that has driven research for the last decade.

The router graph has a natural clustering into Autonomous Systems (ASes), which are sets of routers under the same management. Producing an accurate picture of the *AS graph* is an important step towards 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 a recent paper [14].

The present paper focuses on traceroute sampling, an approach applicable to the router graph as well as the AS graph. Traceroute sampling 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 ASes of degree k is proportional to $k^{-\alpha}$ for a wide range of k values)

A. Bonato and F.R.K. Chung (Eds.): WAW 2007, LNCS 4863, pp. 1–15, 2007.

[7]. This caused a shift in simulation methodology for evaluating network algorithms and also contributed to the avalanche of recently developed network models which produce power-law degree distributions.

However, the true nature of the AS-graph degree distribution was called into question by computer experiments on synthetic graphs [12,17]. These experiments show that if the sets of monitor and target nodes are too small then traceroute sampling will produce a power-law degree distribution, even when the underlying graph has a tightly concentrated degree distribution. Theoretical follow-up work proved rigorously that in many non-power-law graphs, including random regular graphs, an idealized model of traceroute sampling yields power-law degree distributions [4,1].

Subsequent computer experiments have led some to believe that the bias inherent to traceroute sampling can be ignored, at least for making a qualitative distinction between scale-free and homogeneous graphs, when using a large enough set of monitor nodes [9]. This is also supported by an analysis using the statistical physics technique of mean field approximation [5].

1.1 Our Contribution

This paper proposes a new way forward in the struggle to characterize the degree distribution of the AS graph. Our contribution has three parts:

1. We derive a statistical technique for reducing the bias in traceroute sampling;
2. We verify the technique experimentally and theoretically, in the framework previously studied in [12,4];
3. We use the traceroute bias-reduction technique to generate a more accurate picture of the AS degree distribution over time, which suggests that aspects of commercially available technology are reflected in the network topology.

Our approach for reducing the bias in traceroute sampling is based on a technique from biostatistics, the multiple-recapture census, which has been developed for estimating the size of an animal population [18] (this technique also has applications to proofreading [19]). However, we do not have the benefit of independent random variables which are central to the animal counting and proofreading statistics, and so we must adapt the technique to apply to random variables with complicated dependencies.

To provide some evidence that this bias-reduction technique actually reduces bias, we consider a widely used model of traceroute sampling, which assumes that data travels from monitor to target along the shortest path in the network. It is generally believed that the path that data actually takes is *not* the shortest path, but that the shortest path is an acceptable approximation of the actual path (see [13] for a discussion of this approximation). In this model, it is possible to check theoretically and experimentally that the bias reduction provides a better estimate of the degree distribution. We show that the new estimation is asymptotically unbiased for the Erdős-Rényi random graph $G_{n,p}$ when $np \gg \log n$, and that it gives improved estimates for finite instances from a variety of different graphs.

Finally, we use the bias-reduction technique on real data, traceroute samples from the internet. The new estimate of the AS-graph degree distribution is still scale-free over two orders of magnitude, with an exponent very similar to the uncorrected degree distribution (see Figure 1). A by-product of bias reduction is the removal of all vertices with degree less than 3, and this increases the average degree. For example, in March 2004 (the month used for comparison in [14]), the biased estimate of average degree is 6.29, while after bias reduction the average degree is 12.66 (which is very close to 12.52, the biased average degree when restricted to vertices of degree at least 3). An interesting feature in the bias-reduced AS degree distribution (from March 2004) is the lack of nodes with degree between 65 and 90; at the time, a popular router maker offered a router which provided up to 64 ports per chassis. In March 2002, before this product was available, there was no dearth of 65 degree nodes.

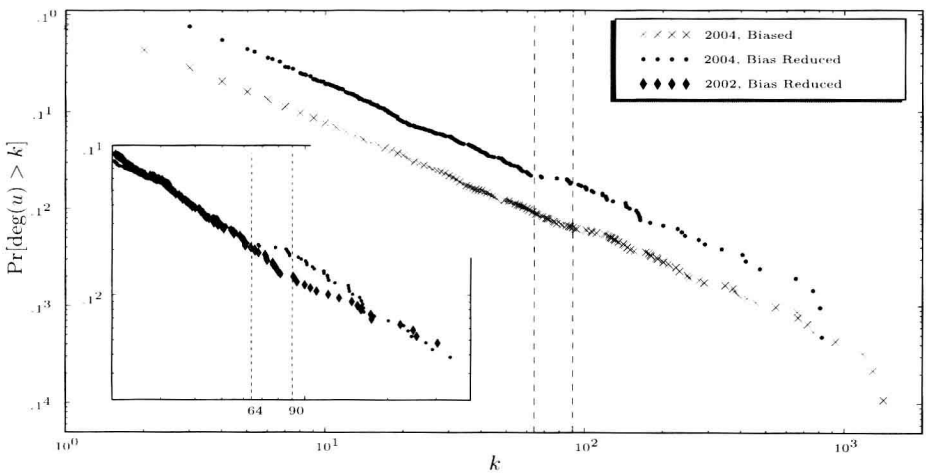

Fig. 1. Degree sequence ccdf estimates for the AS graph (from CAIDA skitter). Main panel: March, 2004, with and without bias reduction. Inset: a portion of ccdf for March, 2004 and March, 2002, both with bias reduction. The nodes with degree between 65 and 90 in 2002 have disappeared in 2004.

1.2 Related Work

Internet mapping by traceroute sampling was pioneered by Pansiot and Grad in [15], and the scale-free nature of the degree distribution was observed by Faloutsos, Faloutsos, and Faloutsos in [7]. Since 1998, the Cooperative Association for Internet Data Analysis (CAIDA) project *skitter* has archived traceroute data that is collected daily [10]. The bias introduced by traceroute sampling was identified in computer experiments by Lakhina, Byers, Crovella, and Xie in [12] and Petermann and De Los Rios [17], and formally proven to hold in a model of one-monitor, all-target traceroute sample by Clauset and Moore

[4] and, in further generality, by Achlioptas, Clauset, Kempe, and Moore [1]. Computer experiments by to Guillaume, Latapy, and Magoni [9] and an analysis using the mean field approximation of statistical physics due to Dall'Asta, Alvarez-Hamelin, Barrat, Vázquez, and Vespignani [5] argue that, despite the bias introduced by traceroute sampling, some sort of scale-free behavior can be inferred from the union of traceroute-sampled paths.

The present paper provides a new avenue for investigating these controversial questions, by developing a method for *correcting* the bias introduced by traceroute sampling. Another recent paper by Viger, Barrat, Dall'Asta, Zhang and Kolaczyk applied techniques from statistics to reduce the bias of traceroute sampling [21]. That paper focused on estimating the number of nodes in the AS graph, and applied techniques from a different problem in biostatistics, estimating the number of species in a bioregion. The problem of correcting bias in sampled networks has a long history in sociology, although the biases in that domain seem somewhat different; see the surveys by Frank, by Klovdahl, or by Salganik and Heckathorn for an overview [8,11,20].

In addition to traceroute sampling, maps of the AS graph have been generated in two different ways, using BGP tables and using the `WHOIS` database. A recent paper by Mahadevan, Krioukov, Fomenkov, Dimitropoulos, claffy, and Vahdat provides a detailed comparison of the graphs that result from each of these measurement techniques [14].

1.3 Outline of What Follows

The new estimator for the degree of a node in the AS graph is developed from multiple-recapture population estimation in Section 2. Section 3 argues that this estimator generates an asymptotically unbiased degree distribution for the Erdős-Rényi graph $G_{n,p}$ when $p \gg \log n$, which rigorously demonstrates that the new estimator can reject a null hypothesis. Section 4 presents additional evidence that the new estimator reduces the bias of traceroute sampling, in the form of computer experiments on synthetic networks. Section 5 provides a comparison between the degree sequence predicted by the new estimator and the previous technique, and details how, after bias reduction, the degree distribution may reflect economic and technological factors present in the system, i.e., there a significantly larger marginal cost of adding a 65th neighbor than adding a 64th neighbor when using the Juniper T320 edge router. Section 6 provides a conclusion and focuses on directions of future research to strengthen this approach.

2 Estimation Technique

The classical capture-recapture approach to estimating an animal population has two phases. First, an experimenter captures animals for a given time period, marks them (with tags or bands), and releases them, recording the total number of animals captured. Then, the experimenter captures animals for a second time period, and records both the number of animals recaptured and the total number

of animals captured during the second period. If A denotes the number of animals captured in phase one, B denotes the number captured during phase two, and C denotes the number captured in phase one and captured again in phase two, then an estimate of total population size is given by

$$\widehat{N} = \begin{cases} \frac{AB}{C}, & \text{if } C \neq 0; \\ \infty, & \text{otherwise.} \end{cases}$$

If the true population size is N, and each animal is captured or not captured during each phase independently, with probability p_1 during phase one and probability p_2 during phase two, then $\widehat{N}$ is the maximum likelihood estimate of N [18]. For more than two phases, the maximum likelihood estimator does not have a simple closed form, but it can be computed efficiently using the techniques developed in [18].

When estimating the degree of a particular AS by traceroute sampling, each edge corresponds to an animal, and each monitor node corresponds to a recapture phase. Unfortunately, in this setting there is no reason to believe that the events "monitor i observes edge j" are independent. Indeed, when shortest-path routing is used (as an approximation of BGP routing), these events are highly dependent. However, it is still possible adapt the capture-recapture estimate to reduce bias in this case.

Let G be a graph, and let s and t be monitor nodes in G. Let G_s be the union of all routes discovered when sending packets from s to every node in the target set. Define G_t analogously. Let $N_s(u)$ denote the neighbors of u in G_s and define $N_t(u)$ analogously.

Using this notation, the modification of the capture-recapture estimate proposed for traceroute sampling is given by

$$\widehat{\deg}_{s,t}(u) = \begin{cases} \frac{|N_s(u)| \cdot |N_t(u)|}{|N_s(u) \cap N_t(u)|}, & \text{if } |N_s(u) \cap N_t(u)| > 2; \\ \infty, & \text{otherwise.} \end{cases}$$

When more than 2 monitor nodes are available, pair up the monitors, consider the estimates given by each pair that are not ∞, and for the final estimator, use the median of these values. So, if the monitor nodes are paired up as $(s_1, t_1), (s_2, t_2), ..., (s_k, t_k)$ then

$$\widehat{\deg}(u) = \text{median}\left(\left\{\widehat{\deg}_{s_i,t_i}(u) \neq \infty\right\}\right).$$

This degree estimator can also provide an estimate of the cdf of the degree distribution (i.e., the fraction of nodes with degree at most k) according to the formula

$$\widehat{d_{\leq k}} = \widehat{\Pr}[\deg(u) \leq k] = \frac{\#\{u : \widehat{\deg}_{s,t}(u) \leq k\}}{\#\{u : \widehat{\deg}_{s,t}(u) < \infty\}}.$$

Discussion: It may seem wasteful to consider the median-of-two-monitors estimate instead of combining all available monitors in a more holistic manner.

However, we have conducted computer experiments with maximum likelihood estimators for multiple-recapture population estimation with more than two phases, and the adaptations we have considered thus far perform significantly worse than the median-of-two-monitors approach above. This is probably due to the complicated dependencies of several overlapping shortest-path trees. However, the exploration we have conducted to date is not exhaustive, and does not rule out the possibility that a significantly better estimator exists.

3 Theoretical Analysis

This section and the next intend to provide some assurance that repeated application of $\widehat{\deg}(u)$ is an accurate way to estimate the degree distribution of the sampled graph.

This section provides a theoretical analysis of the performance of $\widehat{\deg}(u)$ in a very specific setting: when the underlying graph is the Erdős-Rényi graph $G_{n,p}$ with n sufficiently large, $np \gg \log n$, and every vertex is a target node. For the purpose of analysis, this section and the next assume that traceroute finds a shortest path from monitor to target. This is the same setting that is considered in [4].

Theorem 1. *Let $G \sim G_{n,p}$ be a random graph with $np = d \gg \log n$, and let s, t, and u be uniformly random vertices of G. Then, for any k, with high probability,*

$$\widehat{d_{\leq k}} = \frac{\#\{u : \widehat{\deg}(u) \leq k\}}{\#\{u : \widehat{\deg}(u) < \infty\}} = \frac{\#\{u : \deg(u) \leq k\}}{n} \pm \mathcal{O}(1/d)\,.$$

Proof sketch: The analysis *two* breadth-first-search trees in a random graph is difficult when the average degree is small. But, for d moderately large, as in this theorem, the situation is simpler.

It follows from the branching-process approximation of breadth-first search that with high probability there are $(1 \pm \epsilon)d^i$ vertices at distance exactly i from s (or t) when $i < (\log n)/(\log d)$. Thus, almost all vertices are distance $\lceil (\log n)/(\log d) \rceil$ apart. For ease of analysis, suppose that $\ell = (\log n)/(\log d)$ is an integer.

So, with high probability, if u is at distance ℓ from s or t then it is a leaf node in G_s or G_t. In this case, $|N_s(u) \cap N_t(u)| \leq 1$ and therefore $\widehat{\deg}(u) = \infty$.

Now, consider the case where vertex u is distance i from s and distance j from t, where $i, j < \ell$. Let $N(u)$ denote the neighbors of u in G, and then let S be the set of vertices within distance i of s in G and let T be the set of vertices within distance j of t in G. Conditioned on S, T and $N(u)$, the set of indicator random variables

$$\Big\{ \mathbf{1}[v \in N_s(u)], \mathbf{1}[v \in N_t(u)] \colon v \in N(u) \setminus (S \cup T) \Big\}$$

is independent, and, for $v \in N(u) \setminus (S \cup T)$, $\Pr[v \in N_s(u)]$ and $\Pr[v \in N_t(u)]$ are functions of S and T, but constants with respect to v, i.e., $\Pr[v \in N_s] = p_s$