A Simulation Study of Network Discovery Strategies*

Felix Eberhard, Alexander Hall†
ETH Zürich

Thomas Erlebach‡
University of Leicester

Abstract

Due to its fast, dynamic, and distributed growth process, it is hard to obtain an accurate map of the Internet. In many cases such a map—representing the structure of the Internet as a graph with nodes and links—is a prerequisite when investigating properties of the Internet. A common way to obtain such maps is to make certain local measurements at a small subset of the nodes and then to combine these in order to "discover" (an approximation of) the actual graph. Each of these measurements is potentially quite costly. It is thus a natural objective to minimize the number of measurements which still discover the whole graph. We consider this problem for a specific type of measurements and compare four simple greedy strategies in an experimental analysis. Our results show that one can discover accurate information about the structure of large and complex networks using a surprisingly small number of queries.

KEYWORDS: Internet discovery, simulation experiments, complex networks, random graphs.

1 INTRODUCTION

An important aspect in the study of complex networks is the methodology that is used to obtain information about the nodes and links of an unknown network. Before the structure of a network can be analyzed and interpreted, one needs to measure the network in order to discover its nodes and links. In many cases, measuring the network is a nontrivial task, and obtaining accurate and complete information is a challenging problem. The questions arising include: What type of measurements should be carried out? How many measurements are needed for different types of networks? What is the best strategy for minimizing the number of measurements needed to get an accurate reconstruction of the network?

We are interested in the problem of discovering the presence and absence of links in an unknown communication network. The prime example of a communication network whose structure is difficult to determine is the Internet. Owing to its large scale and distributed growth, there is no easy method to obtain an accurate and complete map of the Internet. We consider a model of network discovery in which the set of nodes of the network is known in advance and where a measurement at a node v yields the set of all edges on shortest paths between v and any other node of the graph. We refer to a measurement carried out at a node v as a query at v. This model is motivated by approaches to discovering the Internet (on the router level or on the level of autonomous systems) that are based on traceroute experiments [1, 2] from selected sources or on the analysis of BGP routing tables at

*Work partially supported by European Commission - Fet Open project DELIS IST-001907 "Dynamically Evolving Large Scale Information Systems."

†E-mail: alex.hall@gmail.com

‡Corresponding author. Department of Computer Science, University of Leicester, University Road, Leicester LE1 7RH, UK. Phone: +44-(0)116-2523411. Fax: +44-(0)116-2523604. E-mail: t.erlebach@mcs.le.ac.uk

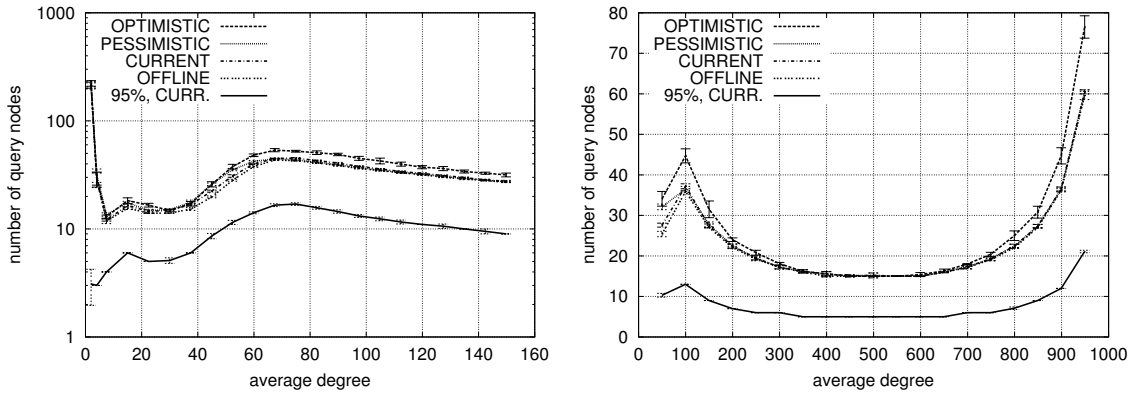

Figure 1: $G_{n,p}$: Erdős-Rényi random graphs on $n = 1{,}000$ nodes. Between a pair of nodes an edge is present with probability p. This parameter is varied. In the charts the expected degree $p \cdot (n-1)$ of a node is given on the abscissa.

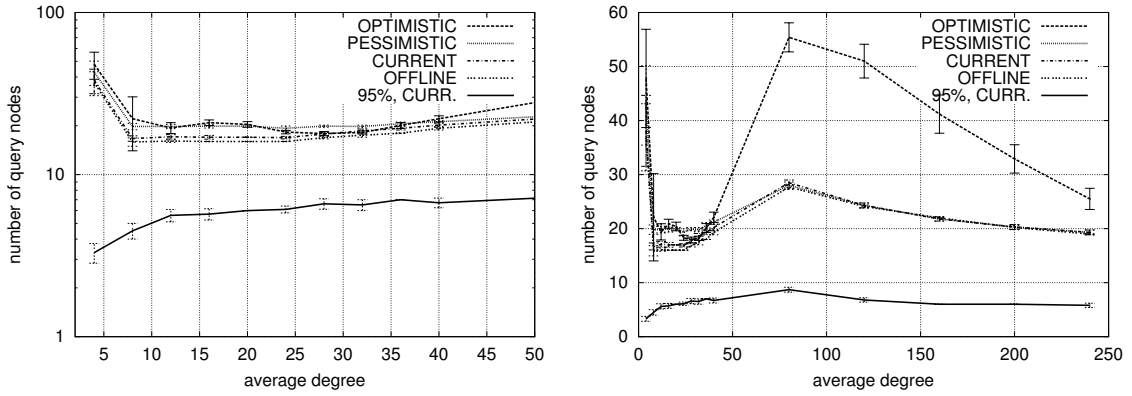

Figure 2: Barabási-Albert random graphs. The average degree of a node is given on the abscissa.

selected BGP routers [3]. A query at v in our model roughly corresponds to carrying out traceroute experiments at v, or obtaining the BGP routing table at v. We implement several discovery strategies and study experimentally the number of queries they need to discover different types of networks.

2 MODEL AND SIMULATION SETUP

Following [4], we model the network discovery problem as follows. The network is represented as an undirected graph with n nodes. A link between two nodes of the network corresponds to an edge between these two nodes in the graph. If there is no link between two nodes u and v, we say that there is a non-edge between u and v. A query at a node v discovers all edges and non-edges whose endpoints have different distance from v. A network discovery strategy makes queries until all edges and non-edges have been discovered. For the selection of the next query, the results of all previous queries can be taken into account. The goal is to use as few queries as possible to discover all edges and non-edges. The difficulty in selecting good queries arises from the fact that the amount of information discovered by a query may depend on the parts of the network structure that are still unknown.

We have implemented several strategies for selecting the next query vertex: Strategy CURRENT selects the next query so as to maximize the number of newly discovered non-edges under the as-

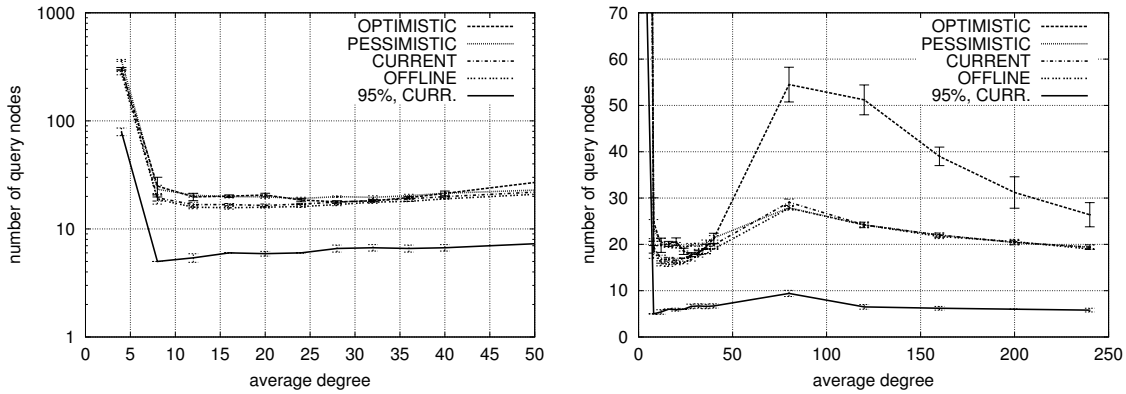

Figure 3: Dorogovtsev-Mendes-Samukhin random graphs. The average degree of a node is given on the abscissa.

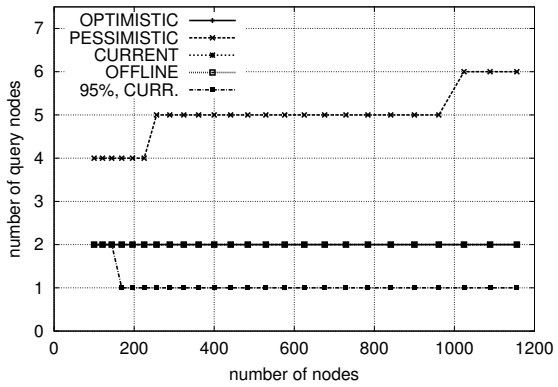

Figure 4: Grids of increasing sizes from 10×10 to 34×34. The number of nodes is given on the abscissa.

sumption that all edges have already been discovered. Strategies OPTIMISTIC and PESSIMISTIC compute upper and lower bounds on the number of newly discovered edges and non-edges at each potential query vertex and select the query that maximizes the upper or lower bound, respectively. For comparison, we have also implemented a strategy OFFLINE that knows the network and computes, using a greedy set-cover heuristic, a small set of queries to discover the whole network. Since OFFLINE has full information, it can be expected to perform better than the three other strategies, and we thus employ it as a benchmark strategy.

In our simulation experiments, we use various types of network graphs (Erdős-Rényi random graphs [5], Barabási-Albert scale-free random graphs [6], Dorogovtsev-Mendes-Samukhin scale-free random graphs [7], and grid graphs) and run the implemented discovery strategies on them.

3 RESULTS

First, we compare the different strategies with each other concerning the number of queries required to discover the whole graph in each of the different random graph models. We generate graphs with $n = 1{,}000$ nodes and varying average degree, and we record the average number of queries that each of the strategies requires. The results are shown in Figures 1–3. The charts on the left-hand side cover smaller degrees and use a logarithmic scale on the vertical axis, whilst those on the right-hand side

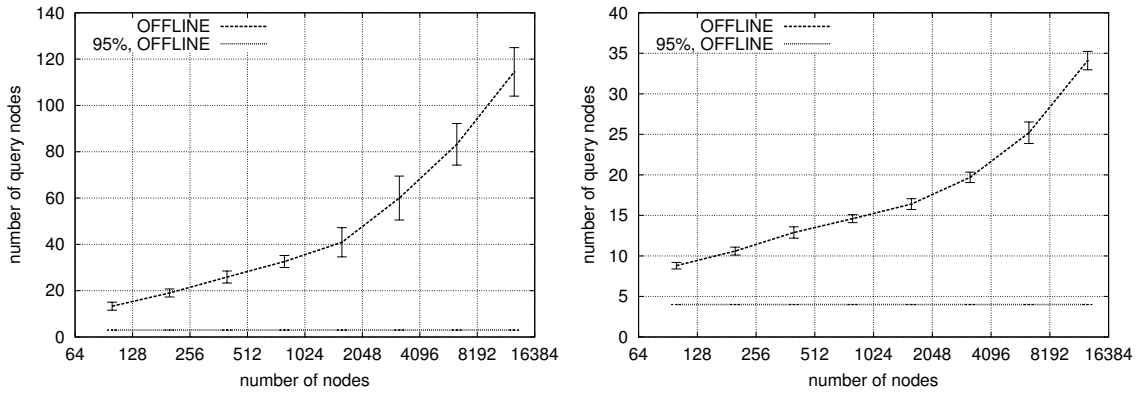

Figure 5: Barabási-Albert random graphs. The average degree is fixed to 4 for the left chart and to 10 for the right chart. The number of nodes is increased in exponential steps.

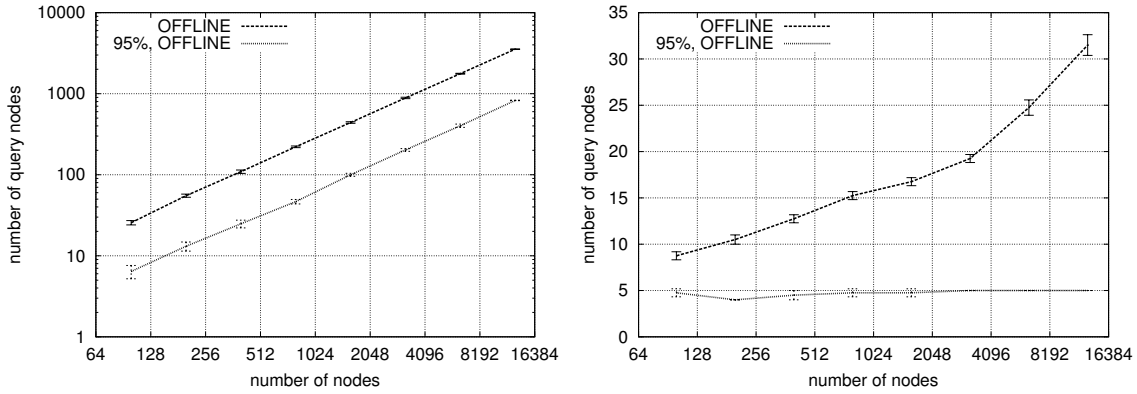

Figure 6: Dorogovtsev-Mendes-Samukhin random graphs. The average degree is fixed to 4 for the left chart and to 12 for the right chart. The number of nodes is increased in exponential steps.

cover a larger range of degrees and use a linear scale. We find that the number of queries required by PESSIMISTIC, CURRENT and OFFLINE is roughly the same and slightly smaller than that of OPTIMISTIC. This indicates that a good selection of queries can be made in spite of the initial lack of information about the graph structure. We also find that the absolute number of queries required to discover a network is often surprisingly small. For networks with 1,000 nodes and average degree ranging from 10 to 250, only 10–60 queries are usually sufficient to discover the whole network. This holds for all three types of random graphs.

Grid graphs were the only graphs for which OPTIMISTIC outperformed PESSIMISTIC (see Figure 4; here, we have varied the number of nodes). This is due to the special property of grids that two queries (in two adjacent corners) are sufficient to discover the whole graph, independent of the number of nodes.

In the charts of Figures 1–4 we also show the number of queries required by CURRENT in order to discover 95% of the edges and 95% of the non-edges (labeled "95%, CURR."). It is a striking observation that more than 10 queries are rarely needed for this purpose.

We have also studied how the number of queries required to discover scale-free random graphs changes if the size of the network grows. The results for strategy OFFLINE are shown in Figures 5 and 6. Each of the charts shows the number of queries needed for networks with increasing number of nodes and fixed average degree. As expected, one finds that when the number of nodes is increased,

the number of queries required to discover the network increases as well. However, when checking how many queries suffice to discover 95% of the edges and 95% of the non-edges of a graph (labeled "95%, OFFLINE"), we notice that for Barabási-Albert graphs this number appears to be a constant smaller than 10 as the number of nodes varies from 100 to 12,800. This indicates that extremely few queries can be sufficient to discover a huge portion of an unknown network, independent of the size of the network. For Dorogovtsev-Mendes-Samukhin random graphs with average degree 4, however, the number of queries for 95% discovery grows linearly with the number of nodes. Only when the average degree is larger (achieved by varying the generation process so that each new node is made adjacent to the endpoints of several randomly chosen edges), the number of queries required for 95% discovery appears to be a constant for this model as well; see the chart on the right-hand side of Figure 6, where the average degree is fixed to 12.

4 CONCLUSION

Our simulation results show that, under the query model considered, it is possible to discover accurate information about the structure of large and complex networks using a moderately small number of queries. If the goal is to discover 95% of the edges and 95% of the non-edges, it even appears that a constant number of queries is often sufficient, independent of the network size. In future work, we would like to investigate this effect from a more theoretical perspective. Furthermore, we plan to study the influence of weaker query models on the number of queries required to discover a network.

References

[1] Dall'Asta L, Alvarez-Hamelin I, Barrat A, Vázquez A, Vespignani A: A statistical approach to the traceroute-like exploration of networks: theory and simulations. *Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN'04)*, LNCS 3405, Springer, pages 140–153, 2005.

[2] DIMES Project. http://www.netdimes.org/

[3] University of Oregon Route Views Project. http://www.routeviews.org/

[4] Beerliova Z, Eberhard F, Erlebach T, Hall A, Hoffmann M, Mihaľák M, Ram LS: Network Discovery and Verification. *31st International Workshop on Graph-Theoretic Concepts in Computer Science (WG'05)*, LNCS, Springer, 2005. To appear.

[5] Erdős P, Rényi A: On Random Graphs I, *Publ. Math. Debrecen*, 6:290–297, 1959.

[6] Barabási AL, Albert R: Emergence of scaling in random networks, *Science*, 286:509–512, 1999.

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