Brief Announcement: On the Hardness of Topology Inference

H.B. Acharya[1] and Mohamed Gouda[1,2]

[1] Department of Computer Science
University of Texas at Austin
[2] National Science Foundation
{acharya,gouda}@cs.utexas.edu

Abstract. Many systems require information about the topology of networks on the Internet, for purposes like management, efficiency, testing of new protocols and so on. However, ISPs usually do not share the actual topology maps with outsiders. Consequently, many systems have been devised to reconstruct the topology of networks on the Internet from publicly observable data. Such systems rely on traceroute to provide path information, and attempt to compute the network topology from these paths. However, traceroute has the problem that some routers refuse to reveal their addresses, and appear as anonymous nodes in traces. Previous research on the problem of topology inference with anonymous nodes has demonstrated that it is at best NP-complete. We prove a stronger result. There exists no algorithm that, given an arbitrary trace set with anonymous nodes, can determine the topology of the network that generated the trace set. Even the weak version of the problem, which allows an algorithm to output a "small" set of topologies such that the correct topology is included in the solution set, is not solvable: there exist trace sets such that any algorithm guaranteed to output the correct topology outputs at least an exponential number of networks. We show how to construct such a pathological case even when the network is known to have exactly two anonymous nodes.

1 Introduction

Several systems have been developed to reconstruct the topology of networks in the Internet from publicly available data - [4], [3]. In such systems, Traceroute [2] is executed on a node, called the source, by specifying the address of a destination node. This execution produces a sequence of identifiers, called a *trace*, corresponding to the route taken by packets traveling from the source to the destination. For example, a trace may be (a, b, c, d, e) where a, b, c, d, e are the unique identifiers (IP addresses) of nodes in the network. We assume a trace is undirected, that is, traces (a, b) and (b, a) are equivalent.

A trace set T is generated by repeatedly executing Traceroute over a network N, varying the source and destination. The problem of reconstructing the topology of the network which generated a trace set, given the trace set, is the *network tracing problem.*

S. Dolev et al. (Eds.): SSS 2010, LNCS 6366, pp. 285–287, 2010.

In our earlier work [1], we studied network tracing for the special case where no node is consistently anonymous, but nodes may be *irregular* - anonymous in some traces, but not in others. In this paper, we extend the theory of network tracing to networks with consistently anonymous nodes.

Clearly, a trace set is generable from a network only if every trace in the trace set corresponds to a path in the network. However, if we define this as sufficient, it is trivial to see that a trace set is generable from many different networks. For example, $T = \{(a, *_1, b), (a, *_2, b)\}$ is generable from any network with nodes a, b, and at least one anonymous node between a and b.

To mitigate this problem, we add two conditions:

- Complete coverage. Every node and every edge must appear in the trace set.
- Consistent routing. For every two distinct nodes x and y, if x and y occur in two or more traces in T, then the exact same set of nodes occur between x and y in every trace in T where both x and y occur.

However, if we modify the trace set slightly to $T' = \{(a, *_1, b), (a, *_2, c)\}$, we see that T' can still be generated from two networks - one where $*_1$ and $*_2$ are the same anonymous node, and one where they are distinct. To solve this problem, we add the assumption that the network that generated a trace set is *minimal*: it is the smallest network, measured by node count, from which the trace set is generable. This forces $*_1$ and $*_2$ to be the same node.

Clearly, these are strong assumptions; it may be argued that in practical cases, both consistent routing and the assumption of minimality may fail. However, even under these assumptions, is the network tracing problem solvable?

Unfortunately, we show in the following section that the answer is negative.

2 The Hardness of Network Tracing

We demonstrate how to construct a trace set which is generable from an exponential number of networks with two anonymous nodes, and no networks with one or fewer anonymous nodes.

It is of interest to note that our results are derived under a network model with multiple strong assumptions (stable and symmetric routing, unique identifiers, and complete coverage). As no algorithm can solve the minimal network tracing problem, even under the friendly conditions of our model, we conclude the problem resists the strongest known network tracing techniques (such as Paris Traceroute to detect artifact paths, and inference of missing links).

We begin by constructing a very simple trace set with only one trace,

$$T_{0,0} = \{(a, *_1, b_1)\}$$

Next, we introduce a new b-node b_2, which is connected to a through an anonymous node $*_2$. To ensure that $*_2$ is not a previously seen anonymous node, we add the trace $(b_1, *_3, a, *_4, b_2)$. By consistent routing, $*_1 = *_3$ and $*_2 = *_4$, but $*_1 \neq *_2$. (Note that consistent routing precludes routing loops. As $*_3$ and $*_4$ occur in the same trace, they cannot be the same node.)

$$T_{1,0} = \{(a, *_1, b_1), (a, *_2, b_2), (b_1, *_3, a, *_4, b_2)\}$$

We now define operation "$Op2$". In $Op2$, we introduce a new non-anonymous node (c_i). We add traces such that c_i is connected to a through an anonymous node, and is directly connected to all b and c nodes.

A single application of $Op2$ to the trace set $T_{1,0}$ produces the trace set $T_{1,1}$ given below.

$$T_{1,1} = \{(a, *_1, b_1), (a, *_2, b_2), (b_1, *_3, a, *_4, b_2), \\ (a, *_5, c_1), (b_1, c_1), (b_2, c_1)\}$$

Now, we apply $Op2$ l times. Every time we apply $Op2$, we get a new anonymous identifier. This new identifier can correspond to a new node or to a previously-seen anonymous node. As we are considering only minimal networks, we know that this is a previously-seen anonymous node. But there are 2 such nodes to choose from ($*_1$ and $*_2$), and no information in the trace set to decide which one to choose. Furthermore, each of these nodes is connected to a different (non-anonymous) b-node, and is therefore distinct from all other anonymous nodes; in other words, each choice will produce a distinct topology.

Hence the number of minimal networks from which the trace set $T_{1,l}$ is generable, is 2^l. As the total number of nodes in a minimal network is n, we also have $n = l + 4$. Thus the number of networks from which $T_{1,l}$ is generable, is exponential in n. An algorithm given this trace set must necessarily output this exponential-sized solution set to ensure it reports the correct topology. Exactly which topology actually generated the trace set cannot be determined.

References

1. Acharya, H.B., Gouda, M.G.: A theory of network tracing. In: 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (November 2009)
2. Cheswick, B., Burch, H., Branigan, S.: Mapping and visualizing the internet. In: Proceedings of the USENIX Annual Technical Conference, Berkeley, CA, USA, pp. 1–12. USENIX Association (2000)
3. Jin, X., Yiu, W.-P.K., Chan, S.-H.G., Wang, Y.: Network topology inference based on end-to-end measurements. IEEE Journal on Selected Areas in Communications 24(12), 2182–2195 (2006)
4. Yao, B., Viswanathan, R., Chang, F., Waddington, D.: Topology inference in the presence of anonymous routers. In: Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies, March 3, vol. 1, pp. 353–363. IEEE, Los Alamitos (April 2003)