Network Tomography and the Identification of Shared Infrastructure

Michael Rabbat
Rice University,
Houston, TX

Robert Nowak
Rice University,
Houston, TX

Mark Coates
McGill University,
Montreal, Quebec

Abstract

This paper considers the problem of identifying network infrastructure that is shared by a collection of end-hosts. This identification is valuable for assessment and design of content distribution systems as well as network performance estimation and simulator design. The network routes connecting a set of sources to a set of receivers form a directed graph. This paper considers the identification of subgraphs shared by two or more sources. We take a system identification approach to the shared subgraph problem, comparing source inputs with receiver outputs. Sets of receivers are then associated with shared subgraphs using this novel multiple source probing scheme. Our methodology does not rely on special-purpose cooperation from internal network elements, and only requires end-to-end measurements that are easy to make. Experiments conducted on a local area network and the Internet demonstrate the potential of our approach.

1. Introduction and Background

Inferring internal characteristics (delays, losses, routes) of wired communication networks from end-to-end measurements is an important type of inverse problem arising in networking. The network identification problem is often called *network tomography* because it is somewhat analogous to the tomography problem arising in medical imaging. Network tomography can entail the estimation of loss rates or delays on internal network links, the estimation of point-to-point traffic flow rates through a network, the identification of network topology and routing, and the inference of other internal characteristics relevant to the performance of a network or inter-network [1].

The goal of this work is to determine what portions of a network infrastructure are common between routes from different sets of senders and receivers based on easily made measurements of the network. The routing topologies can be expressed as graphs, and therefore we call this problem *the identification of shared subgraphs.* This identification is crucial for many purposes:

Simulation: Accurate knowledge of how the networks and inter-networks are connected is important for the purposes of generating accurate models for simulation. It is generally difficult, especially for third parties, to determine this connectivity.

Distribution: Content distribution systems often have many mirror sites, and it is advantageous to place these sites in a manner which ensures that the content is accessible even if certain portions of the Internet become congested or fail. It is also beneficial to assign mirrors to clients in order to optimize the load on each server. Thus, determination of the potential shared topologies between the mirror sites and clients can provide key information about the robustness of the content distribution system.

Estimation: Network tomographic methods designed to estimate the loss rates or delays on internal links rely on accurate information about the routing topologies involved. Determination of shared portions of network topologies is critical to the success of these performance estimation methods.

The approach pursued here is quite distinct from common tools like traceroute, which rely on the cooperation of internal routers to obtain routing information. Instead, our approach assumes no special-purpose cooperation from internal network elements. This is important since already a large proportion of Internet routers do not respond to traceroute requests [2], and this proportion will probably grow in the future. Moreover, switches and other lower-level network elements cannot be queried by traceroute requests; i.e., they are effectively invisible to traditional methods. Nonetheless, such devices can significantly affect network performance. We address these issues by adopting a system identification approach. By actively probing the network and comparing the input from the sources and output at the receivers, we are able to determine which portions of the network infrastructure are shared.

email: {rabbat, nowak}@rice.edu, coates@ece.mcgill.ca. This work was funded by NSF Grant no. MIP-9701692, ARO Grant no. DAAD19-99-1-0349, ONR Grant no. N0014-00-1-0390, and Texas Instruments.

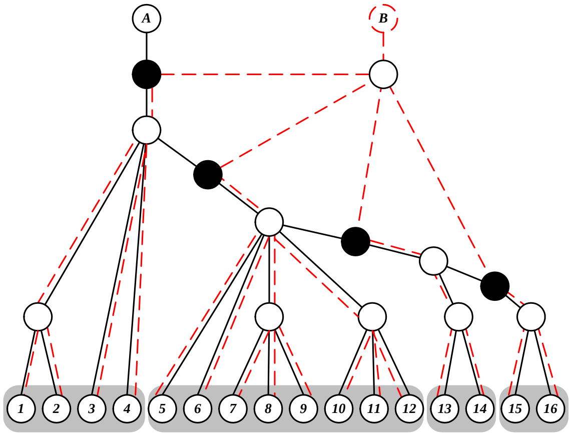

Fig. 1. The routing tree of a local area network. Filled circles represent switches or routers where paths from the sources A and B join, and empty circles represent points where paths branch apart to the different receivers, $1, 2, \ldots, 16$.

2. Shared Infrastructure and Subgraphs

Figure 1 depicts the route graph of a local area network at Rice University. The collection of end-hosts and the routes connecting them can be interpreted as a directed graph, flowing downward from sources A and B to receivers $1, 2, \ldots, 16$. Internal nodes of the graph either represent points where two routes join together or points where two routes branch apart. The problem considered in this paper is to identify shared subgraphs. Shared subgraphs are the trees below shared *joining points* in the paths from A and B to the receivers, and can be associated with the sets of receivers connected to those trees. In the example in Figure 1, there are four shared subgraphs, associated with receiver sets $\{1, \ldots, 4\}$, $\{5, \ldots, 12\}$, $\{13, 14\}$, $\{15, 16\}$. Our goal is to identify these shared subgraphs, or more precisely the corresponding receiver sets, without knowledge of the routing topology. We employ a simple probing method which only requires a minimal level of cooperation between end-hosts (sources and receivers), and does not require any special-purpose cooperation from the internal nodes (e.g., routers/switches). The method relies on the preservation of the ordering of packets as they flow from the sources to the receivers. Packet ordering is preserved in many networks, however our method is not applicable in cases where this ordering may be disrupted (e.g., in load-balancing networks).

Much of the previous work done in this area [1, 3–5] has focused on the identification of logical topologies from a single source to multiple receivers. In such cases the topology takes the form of a tree with the source at the root and the receivers as leaves. The identification of shared subgraphs by the approach described in this paper can be combined with such methods to obtain the graph associated with multiple receivers. In Figure 1, the graph from A to the receivers and the graph from B to the receivers are both trees. Given both of these trees (perhaps estimated by a single sender technique [3–5]), the identified shared subgraph can be used to "merge" the two trees together [6].

The remainder of the paper is organized as follows. In Section 3 we formally state the problem. Our active probing and hypothesis testing framework is discussed in Section 4. In Section 5 we present results from an experiment conducted over a university local area network, and we conclude in Section 6.

3. Problem Statement

In our analysis, the following simplifying assumptions are made. First, we assume that there is a unique path from each source to each receiver. For some cases where the paths between end-hosts are very long, this assumption may not hold. In general it is not an unreasonable assumption, especially when considering campus and local area networks. Our second assumption is that these unique routes are constant over short periods of time. For an experiment conducted over a period of 5 to 10 minutes this is reasonable.

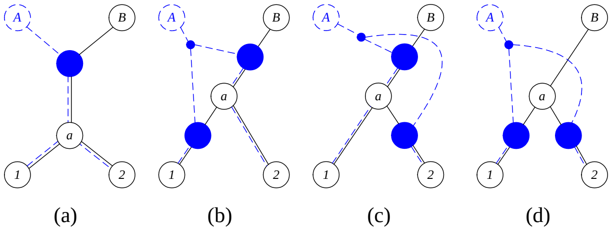

Fig. 2. Four possible entry cases for a two-source, two-receiver network. The filled circles indicate joining points and the empty circles indicate branching points.

Consider paths between the pair of sources, A and B, and the pair of receivers, 1 and 2. Figure 2 depicts examples of possible ways these four end-hosts can be connected. In particular, we are interested in distinguishing the case where the subgraph below internal node a is shared, depicted in Figure 2(a), from the other *unshared* cases Figure 2(b-d).

In developing our technique for determining whether or not there is a shared subgraph in the topology connecting two sources and two receivers, we also adhere to the following design principles. (1) Our technique should not depend on eliciting special responses from network internal devices. We only assume that these devices will route packets between end-hosts. By adopting this principle we avoid the limitations of traceroute-based techniques. (2) The measurement process should be non-intrusive. We wish to restrict the number of packets being actively sent into the network to a bare minimum, so as not to disrupt other traffic on the network. (3) It should be possible to take measurements over a short period of time, so that the probability that routes are stationary over this period is high.

4. Probing to Detect Shared Paths

Our measurement technique involves sending packets from two sources in a coordinated fashion. The driving idea behind our probing methodology is that *the order in which packets arrive at each receiver is the order in which they reach the joining point.* In other words, a delay is incurred along each logical link traversed after packets reach the joining point, but the order in which they reach the joining point remains intact.

The basic probe in our methodology has the following form. Consider the two-source, two-receiver networks depicted in Figure 2. A probe consists of four packets: one packet from A to receiver 1, denoted $p_{A,1}$, one from A to 2, $p_{A,2}$, one from B to 1, $p_{B,1}$, and one from B to 2, $p_{B,2}$. At time t_1 packets $p_{A,1}$ and $p_{B,1}$ are sent. At time $t_1 + \Delta$ packets $p_{A,2}$ and $p_{B,2}$ are sent. For clarity, we first explain how the method applies in the case of constant but unknown delays. If the subgraph from internal node a to 1 and 2 is shared, as in Figure 2(a), then the order in which the packets are received at 1 and 2 should have the following property. If packet $p_{A,1}$ arrives before $p_{B,1}$, then $p_{A,2}$ should arrive before $p_{B,2}$, and vice-versa. This simply reflects the fact that the joining point in the paths from the two sources is the same for both receivers. In this case, we say that the packets are received *in order*. Alternatively, if the joining point is not shared, then the order in which the packets are received may not agree at the two receivers.

However, if the packets arrive in order, one cannot immediately conclude that the subgraph from node a to 1 and 2 is shared. This is because the delays to the joining point(s) may be such that the packets arrive in order. This situation, however, can be ruled out by repeating the probing exercise above with different time offsets. That is, send $p_{A,1}$ and $p_{A,2}$ at times $(t_1, t_1{+}\Delta), (t_2, t_2{+}\Delta), \ldots$ and send $p_{B,1}$ and $p_{B,2}$ at times $(t_1', t_1' + \Delta), (t_2', t_2' + \Delta), \ldots$. Figure 3 shows the probing scheme. If the subgraph is shared, then in all these cases, irrespective of the values of t_i and t_i', the ordering should agree at the two receivers. However, the ordering cannot agree in the unshared cases (Figure 2(b)-(d)) for all possible values of t_i and t_i'. For a certain range of $|t_i - t_i'|$, the differing delays from the sources to the unshared joining points will cause the packets to arrive in different orders at the two receivers.

To test for such a range of probings, we set $t_i' = t_i + u_i$, where $\{u_i\}$ are independent and uniformly distributed on the interval $[-D, D]$. Here D is an upper bound on the source-to-receiver delay (which can be easily determined from round-trip time measurements). If the received order is the same at both receivers in all cases, then we conclude the joining point is shared as in Figure 2(a). If not, then we conclude that it is not shared as in Figure 2(b)-(d).

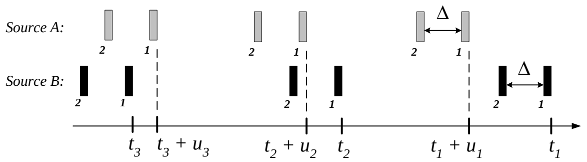

Fig. 3. Packet pairs sent out from each receiver at time t_i. The first packet in each pair is sent to receiver 1, and the second to receiver 2. The spacing Δ is the same for every pair.

In the analysis above, it was assumed that the delays were constant on all links, but in reality there will be some delay variation that could cause mis-ordering. In the shared case, the variation causes a certain percentage ρ of misordered occurrences. The percentage ρ is proportional to σ/D, where σ is the standard deviation of the delay difference. In the unshared cases, the difference between the average delays to the differing joining points, in addition to delay variation, increases the percentage of mis-ordered occurrences significantly.

To gauge the statistical significance of the number of misordered occurrences requires knowledge of σ. It is difficult to estimate σ directly, but a relatively straightforward probing procedure can provide an estimate of ρ. The upper bound is estimated by again sending pairs of probes from each source in the same manner as described above, except that all four packets are sent to one receiver. In such cases, the joining point is always shared (obviously), and thus the number of mis-ordered occurrences in these experiments provides an estimate of ρ.

This entire procedure can be formalized as a binary hypothesis test, and assuming a sufficiently large number of probes one can derive a *Z*-test. This allows us to specify an allowable level α for Type 1 errors (i.e., the probability of declaring an unshared subgraph when in fact the subgraph is shared). For more details see [6]. In all experiments described in this paper, $\alpha = 0.05$.

As stated before, our goal is to identify shared subgraphs. After performing tests of the type described above for each pair of receivers, we have a collection of shared/unshared results for pairs of receivers. We then group receivers together such that members of each group are *mutually shared*. In other words, if the tests for receiver pairs (i, j), (j, k), and (i, k) all turn out shared, then we form the group $\{i, j, k\}$. In the case of the routing topology depicted in Figure 1, this is how we form the final subgraph sets $\{1, \ldots, 4\}$,

$\{5,\ldots,12\}$, $\{13,14\}$, and $\{15,16\}$. Then, given source A's topological tree, the shared subgraph topology consists of all interior nodes whose children are members of a given subgraph set.

Performing tests for pairs of receivers quickly gets expensive, in terms of the number of probes sent, as the number of receivers increases. For N receivers, we can improve the efficiency of our probing algorithm by sending a sequence $\{p_{S,i}\}_{i=1}^{N}$ from each source, S, where the ith packet is directed towards receiver i, and with each packet in the sequence still spaced in time by Δ. After collecting results, the hypothesis tests are still computed pair-wise, but the number of probes, the resource which is more of a concern to us, decreases significantly. In fact, the number of probes required goes from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ with this modification.

5. Experimental Results

We have written a program implementing the ideas discussed above. The program runs under standard Linux/Unix environments and uses the Berkeley sockets API for sending UDP packet probes to the receivers. There are two separate source components and a receiver component. Source A sends packet pairs at regularly spaced intervals in time. Source B controls the experiment and sends packet pairs at random times within each interval, chosen from a uniform distribution. Because the only relevant metric is packet arrival order, no special timing infrastructure is required. The receiver component simply tracks the order in which packets arrive, and sends the results to source B once the experiment has completed. After results have been collected they are processed at source B.

To explore the efficacy of our technique we have conducted experiments in two very different networking environments. The first is a departmental local area network (LAN). The second consists of hosts located at academic and research institutions throughout the United States and Europe. Each scenario presents its own set of difficulties. In the LAN experiment, the delay differences can be very small, making the range of $|t_i - t_i'|$ over which mis-order events occur very small in the unshared case. In the Internet experiment these delay differences are more pronounced, but the delay variance is much higher.

5.1. LAN Experiment

With help from the network management team, we were able to obtain the true physical topology in order to verify our results. Figure 1 depicts the shared subgraph discovered by our algorithm. The results correspond exactly to shared subgraph derived from the true physical underlying network. For the purposes of the experiment, source A's logical tree is assumed to be known, and the points where source B's topology joined were added on to this tree based on the pair-wise receiver test results.

The test bed for this experiment is a subset of hosts from the Rice University ECE departmental local area network (LAN). This environment showcases the strong points of our algorithm. The underlying physical network is composed of a mixture of layer-2 and layer-3 devices, of which both types are detected. The network is contained in a compact area, and as a result packet transmission times are on the order of a few microseconds. Accurate time measurements at this resolution are not easily obtained with standard hardware and operating system configurations.

For this experiment there were 16 receivers with IP addresses on two different subnets. Both subnets reside on the same physical network, which consists of a variety of devices including 3Com SuperStack 3300 and 1000 switches. Note that one variety is a store-and-forward device while the other implements cut-through switching. Our technique finds shared subgraphs regardless of the switching technology implemented at joining or branching points.

Each probe is 68 bytes, including payload, UDP and IP headers. Using 600 microseconds for the random offset bound D is sufficient to encompass the range of possible delays for the short paths of the LAN.

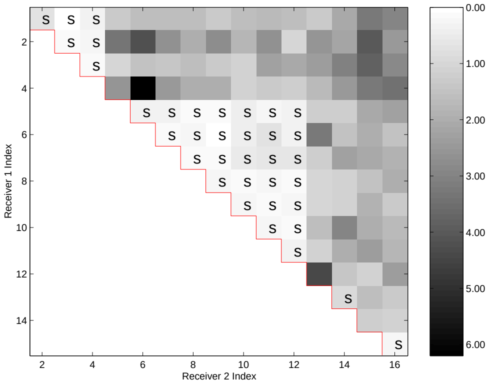

Fig. 4. Experimental results. The x- and y-axes correspond to receivers as labelled in Figure 1. The intensity of the square at position $(i,\ j)$ indicates the observed ratio of mis-ordered events to total measurements for the receiver pair (i, j). If the square at (i, j) is labelled with an ”s”, then the test decided that the paths to receivers i and j share common joining and branching points from the two sources.

In our experiments, all of the decisions were correct in the sense that they agreed with the known logical connectivity. Figure 4 graphically depicts the results of one experiment. We correctly identify the set of shared subgraphs.

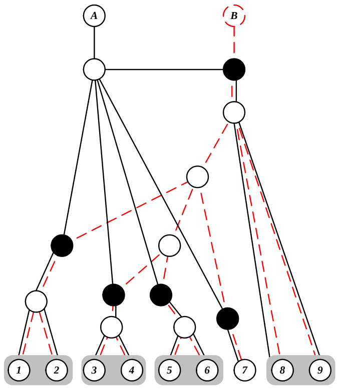

Fig. 5. Logical routing tree of hosts from an experiment conducted over the Internet. Filled circles represent switches or routers where paths from the sources A and B join, and empty circles represent points where paths branch apart to the receivers, $1, 2, \ldots, 9$.

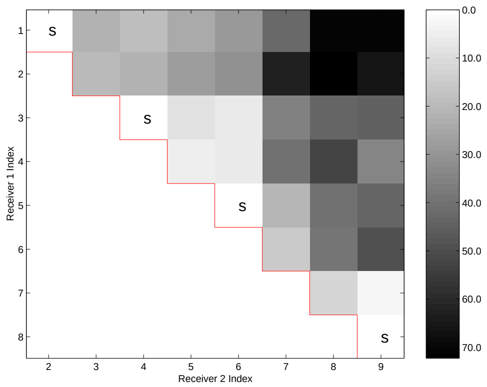

Fig. 6. Experimental results. The x- and y-axes correspond to receivers as labelled in Figure 5. Intensities and markings hold the same meaning as those in Figure 4

5.2. Internet Experiment

In order to explore algorithm performance in an environment very different from the LAN we performed another set of experiments using Internet hosts located in North America and Europe. For these experiments there were 9 receiving hosts located at 5 different academic establishments. The two sources were both situated in North America. Figure 5 shows the logical connectivity between sources and receivers based on measurements made with traceroute.

To adjust for longer delay times we increase the D parameter to 90 milliseconds. In this experiment, we are able to correctly identify pairs of receivers with shared subgraphs. Figure 6 shows the results of the pair-wise experiments. In all shared cases, both hosts in the shared pair were located at the same academic institution.

6. Conclusion

We have developed an active probing framework based on the arrival order of packets at receivers that can be used to determine whether the paths connecting two sources and two receivers are part of a shared subgraph. These shared/unshared results for pairs of receivers are then combined with information about one source's tree topology to obtain a general shared topology. The techniques described are validated through experiments over a university LAN and the Internet.

Work remains to be done in the area of generalizing the probing procedure to more than two sources in order to make it more scalable. We will also explore the development of less intensive multiple source probing methods that monitor for changes in an initially established shared subgraph, perhaps on networks much larger than those considered in our work thus far. We also plan to address issues regarding this work and the fact that the existence of multiple paths from a source to a receiver is always a possibility in the Internet.

7. Acknowledgements

Special thanks go out to Glenn Forbes Fleming Larratt and the IT department at Rice University for their help in validating our experiments over Rice's LAN. We would also like to thank Bobak Nazer for his help in administering the experiments and Rudolf Riedi for his support and insight.

8. References

[1] M. Coates, A. Hero, R. Nowak, and B. Yu, "Internet tomography," *IEEE Sig. Proc. Magazine*, May 2002.

[2] P. Barford, A. Bestavros, J. Byers, and M. Crovella, "On the marginal utility of network topology measurements," in *Proc. IEEE/ACM SIGCOMM Internet Measurement Workshop*, San Fran., CA, November 2001.

[3] A. Bestavros, J. Byers, and K. Harfoush, "Inference and labeling of metric-induced network topologies," Tech. Rep. BUCS-TR-2001-010, Computer Science Department, Boston University, Boston, MA, May 2001.

[4] M.J. Coates, R. Castro, M. Gadhiok, R. King, Y. Tsang, and R.D. Nowak, "Maximum likelihood network topology identification from edge-based unicast measurements," in *Proc. ACM Sigmetrics*, Marina Del Rey, CA, June 2002.

[5] N.G. Duffield, J. Horowitz, F. Lo Presti, and D. Towsley, "Multicast topology inference from mea-

sured end-to-end loss," *IEEE Trans. Info. Theory*, vol. 48, no. 1, pp. 26–45, January 2002.

[6] M. Coates, M. Rabbat, and R. Nowak, "Discovering logical network topologies," Tech. Rep. TREE-0202, Rice University, Houston, TX, November 2002.