Scalable Deterministic Overlay Network Diagnosis

Yao Zhao, Yan Chen, David Bindel
yzhao,ychen@cs.northwestern.edu, dbindel@eecs.berkeley.edu

> "When something breaks in the Internet, the Internet's very decentralized structure makes it hard to figure out what went wrong and even harder to assign responsibility."
> – "Looking Over the Fence at Networks: A Neighbor's View of Networking Research", by Committees on Research Horizons in Networking, National Research Council, 2001.

Internet fault diagnosis is important to end users, overlay network service providers (like Akamai), and Internet service providers (ISPs). For example, with Internet fault diagnosis tools, users can choose more reliable ISPs. However, The modern Internet is heterogeneous and largely unregulated, which renders the Internet diagnosis an increasingly challenging problem.

Though several router-based Internet diagnosis tools have been proposed [1], [2], these tools generally depend on ICMP measurements. ICMP-based tools are subject to ICMP rate limiting, are sensitive to cross-traffic, and are un-scalable. In contrast, many recently-developed tools for *Internet Tomography* use signal processing and statistical approaches to infer link level properties [3], [4], [5], [6] or shared congestion [7] based on end-to-end measurements of IP routing paths. We define *paths* to be IP routing paths between pairs of end hosts; paths are made up of *links*, which are IP connections between routers. The latency along a path is the sum of the latencies along the links that make up the path; and other path properties can similarly be expressed in terms of link properties. The relation between path and link properties can be written as a large linear system; however, as we observed in [8], the linear system is *fundamentally underconstrained*: there exist *unidentifiable links* [9], [8] with properties that cannot be uniquely determined from path measurements. In order to estimate the properties of unidentifiable links, Internet tomography tools often impose statistical assumptions; thus, the accuracy of the predicted link properties is subject to uncertainty in the model assumptions. As shown below, such statistics-based tools are neither *deterministic* nor *scalable*.

Existing tomography systems analyze the temporal correlations among multiple receivers in a multicast-like environment; and with enough probes, they can infer the loss rate of each path segment with high probability. However, their inference results are not *deterministic* or *unique* for two reasons. First, they can only achieve 100% determinism with infinitely many probes. Second, while these systems can obtain very high probability estimates with a certain number of probes (the exact number depends on the depth of the tree and the number of receivers), they suppose an ideal multicast environment. However, given that multicast does not really exist in the Internet, they have to use unicast for approximation. Thus the inference accuracy heavily depends on the cross traffic of the network, and there is no guarantee or bound on the inference accuracy. Furthermore, the iterative refinement algorithms used to compute the link properties are expensive for large networks, and may not always converge. Thus it remains an open problem to find which links or sequences of links can be *uniquely* characterized from end-to-end measurements, for which we will tackle in this paper.

Problem Definition and Solution Here we define the *granularity* as the length of a sequence of links on a path. We would like a fine-grained characterization of the overlay network behavior, *i.e.*, to characterize the properties of very short sequences of links. Fine-grained characterization is important for congestion and failure diagnosis, since the granularity determines how well we can localize problems. Because the linear system relating link properties to path properties is underconstrained even for a very large overlay network [8], we cannot resolve the properties of each link individually. What, then, is the finest granularity we can attain?

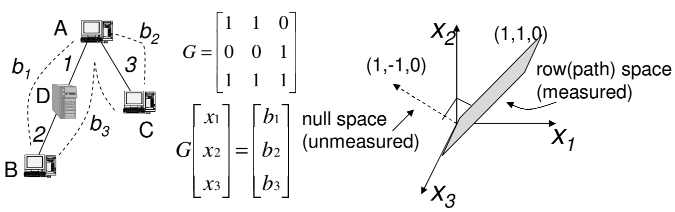


Fig. 1. Sample overlay network.

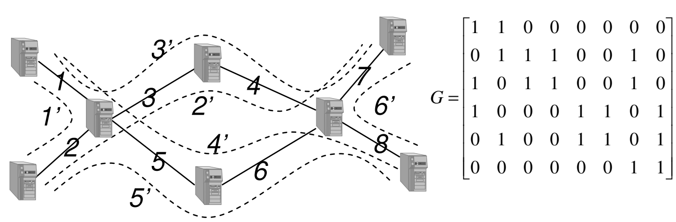
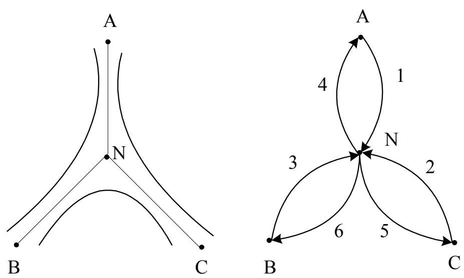


Fig. 2. Sample overlay network with 8 links and 6 paths (*e.g.*, $1'$).

In this paper, we apply an algebraic approach to separate the identifiable and unidentifiable components of each path to address the questions raised before. We define *virtual links* as link sequences of *minimal* length whose properties can be uniquely identified from end-to-end measurements. Fig. 1 shows an example how we use the linear algebraic model to find identifiable virtual links. Here G denotes the path matrix, in which $G_{ij} = 1$ means the jth link is in the ith path. x_i denotes the logarithm of the success ratio of link i and b_i is the logarithm of the success



$$\begin{array}{cc} & \begin{array}{cccccc} \multicolumn{3}{c}{\text{Incoming Links}} & \multicolumn{3}{c}{\text{Outgoing Links}} \\ 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ G = & \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right] \end{array}$$

$\text{Rank}(G) = 5$

Fig. 3. Undirected graph vs Directed graph.

ratio of path i. All identifiable virtual links belong to the row space of G, i.e., a subspace containing all the vectors can be written as linear combination of the row vectors of G. We design efficient algorithms to find the virtual links and infer their loss rates for diagnosis. For example, there are altogether four non-trivial virtual links in Fig. 2, corresponding to link (sequence) 1, 2, $3+4+7$ and $5+6+8$ separately. For example, link 1 is a virtual link, of which the loss rate can be inferred through the linear combination of path success rates as $(b_1 + b_3 - b_2)$.

In contrast to previous statistical inferences, if successful, we will be the first to achieve *deterministic* network tomography, for any network topology, and for both undirected and directed graphs. We call our system a Deterministic Overlay Diagnosis (DOD) system.

To identify the virtual links, we first only use the routing and network topology information. Thus the virtual links are uniquely identified by the inherent path sharing of the Internet. We designed efficient algorithms to find all the virtual links in an undirected graph. Actually, Yuval *et al.* had similar attempt in undirected graph and assumed that it can be extended to directed graph [10]. *However, we found and proved that for a directed graph, each path itself has to be a virtual link, and no path segment can have its property uniquely determined.* For example, Fig. 3 shows a simple star topology in both undirected graph and directed graph. Even the directed graph contains all the 6 end to end paths, any single link in the graph is not identifiable. The situation looks like a deadlock, and all the links can be identified if we can identify one link in advance(*i.e.*breaking the deadlock).

Note that the loss rate measurement can help us to further reduce the granularity of virtual links for diagnosis. In fact, the analysis solely based on topology gives the worst case scenario when all paths are lossy. In practice, there are many good paths with nearly no loss and it is reasonable to say all links in a non-lossy path have no loss. This motivated us to design "good path" algorithms to break the deadlock in directed graph. This effectively solve the problem for diagnosis on directed graphs, and is very useful for one-way congestion/failure location.

The algebraic approach was also used recently for scalable overlay network monitoring to infer the end-to-end path properties [8], but not on the link level. For overlay diagnosis, we naturally inherit the scalability and load balancing from [8]. That is, to diagnose an overlay network of n nodes, we only need to measure $O(n \log n)$ paths instead of all the $O(n^2)$ paths. These load are evenly distributed across the end hosts.

Preliminary Results We evaluated the DOD system through extensive simulations, and further validate our results through Internet experiments. For validation over the Internet, we proposed a novel method of link-level loss rate inference based on IP spoofing. The basic idea is to use IP spoofing to create some new paths that may not exist in normal routings, or it can be viewed as an approximation of source routing.

Both the simulation and Internet experiments give promising results. For the PlanetLab experiments with 135 end hosts (each from different organization), the average diagnosis granularity is only 2.83 hops for 3,714 lossy paths with average path length 15.2. This can be further improved with larger overlay network as shown through our simulation with a real router-level topology from [11]. This suggests we can do very fine-level deterministic diagnosis with reasonable large overlay network.

In addition, the loss rate inference on the virtual links is highly accurate as verified through the cross validation and IP spoof based validation schemes. This is due to the determinism inherent in our diagnosis system. The DOD system is also highly efficient. For the PlanetLab experiments with 135 hosts, the average setup (monitoring path selection) time is 109.3 seconds, and the diagnosis of the 3,714 lossy paths out of 18,090 paths takes only 4.2 seconds on a 2.8GHz P4 machine. For more details and the poster, please refer to `http://list.cs.northwestern.edu/DOD.html`.

References

[1] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson, "User-level internet path diagnosis," in *ACM SOSP*, 2003.

[2] K. Anagnostakis, M. Greenwald, and R. Ryger, "cing: Measuring network-internal delays using only existing infrastructure," in *IEEE INFOCOM*, 2003.

[3] Mark Coates, Alfred Hero, Robert Nowak, and Bin Yu, "Internet Tomography," *IEEE Signal Processing Magazine*, vol. 19, no. 3, pp. 47–65, 2002.

[4] A. Adams et al., "The use of end-to-end multicast measurements for characterizing internal network behavior," in *IEEE Communications*, May, 2000.

[5] T. Bu, N. Duffield, F. Presti, and D. Towsley, "Network tomography on general topologies," in *ACM SIGMETRICS*, 2002.

[6] V. Padmanabhan, L. Qiu, and H. Wang, "Server-based inference of Internet link lossiness," in *IEEE INFOCOM*, 2003.

[7] D. Rubenstein, J. F. Kurose, and D. F. Towsley, "Detecting shared congestion of flows via end-to-end measurement," *ACM Transactions on Networking*, vol. 10, no. 3, 2002.

[8] Y. Chen, D. Bindel, H. Song, and R. H. Katz, "An algebraic approach to practical and scalable overlay network monitoring," in *ACM SIGCOMM*, 2004.

[9] N. Duffield, "Simple network performance tomography," in *ACM IMC*, 2003.

[10] Y. Shavitt, X. Sun, A. Wool, and B. Yener, "Computing the unmeasured: An algebraic approach to Internet mapping," in *IEEE INFOCOM*, 2001.

[11] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet map discovery," in *IEEE INFOCOM*, 2000.