NORTHWESTERN
UNIVERSITY

Electrical Engineering and Computer Science Department

Technical Report
NWU-EECS-09-03
January 12, 2009

What's Wrong with Network Positioning and Where Do We Go From Here?

David R. Choffnes and Fabián E. Bustamante

Abstract

Network positioning systems are essential to good performance in distributed systems that scale to millions of end hosts. Evaluating performance in this environment is particularly challenging. This paper addresses this issue through an empirical study of two alternative classes of network positioning services based on a dataset gathered from more than 43K IP addresses probing over 8M other IPs worldwide. We use more than 1.4 billion network measurements to show that network positioning exhibits noticeably worse performance than previously reported studies. To explain this result, we identify several key properties of this environment that call into question fundamental assumptions driving network positioning research. We close by suggesting a new direction for network positioning that draws from our observations.

Keywords: Network positioning, network coordinates, triangle inequality violations

What's Wrong with Network Positioning and Where Do We Go From Here?

David Choffnes and Fabián E. Bustamante
EECS, Northwestern University
{drchoffnes,fabianb}@cs.northwestern.edu

Abstract

Network positioning systems are essential to good performance in distributed systems that scale to millions of end hosts. Evaluating performance in this environment is particularly challenging. This paper addresses this issue through an empirical study of two alternative classes of network positioning services based on a dataset gathered from more than 43K IP addresses probing over 8M other IPs worldwide. We use more than 1.4 billion network measurements to show that network positioning exhibits noticeably worse performance than previously reported studies. To explain this result, we identify several key properties of this environment that call into question fundamental assumptions driving network positioning research. We close by suggesting a new direction for network positioning that draws from our observations.

1 Introduction

There is a growing number of large-scale distributed systems (e.g., streaming video, VoIP and file sharing [1, 4, 8, 9]) that run on hosts located at the edges of the network (e.g., on desktops or appliances behind NAT boxes on residential links). Because most of these systems are cooperative and deployed at a scale of hundreds of thousands or millions of users, brute-force methods for providing key functionality, such as optimal peer selection, are prohibitively expensive.

In particular, many of these applications could benefit from a scalable way to determine the relative location of hosts in the network. Toward this end, network positioning systems attempt to efficiently estimate the network distance (in terms of latency) between hosts [6, 14, 22]. Because there has been no traces representative of the environment, nor available platform at the *scale* where network positioning is intended to be used, these systems have been commonly implemented and evaluated in simulation and on research testbeds.

Taking advantage of a unique deployment of measurement software deployed on end hosts in a large-scale P2P system, *this paper evaluates how two key classes of network positioning systems fare when deployed at scale and measured in networks where they are used*. To achieve a large, representative dataset, we base our study on latency measurements reported by hosts participating in the Vuze BitTorrent system [20], which provides an operational deployment of Vivaldi [6] and a rich interface for accessing its coordinates. With approximately 1M users online worldwide at any moment, this represents the largest deployment of any network positioning service. Through an extension to this client, currently installed by over 380,000 users, we sample Vivaldi network coordinates and perform network measurements to evaluate its accuracy. We additionally use the latency measurements between hosts to understand Meridian [22] performance in such an environment.

As a result of this largest-ever measurement study of a deployed positioning system, we find that the accuracy of the network coordinate systems is significantly worse from the perspective of end-user clients than when evaluated from the perspective of a research testbed. Next, we show that this inaccuracy leads to significant loss in performance in the case of DHT peer selection. We find that these errors are in part explained by the inherent dimensionality of the worldwide latency space, which is much larger than reported from previous studies that use limited deployments. Finally, using traceroute measurements, we demonstrate that accounting for the Internet's structural properties is essential to accurate distance estimation.

2 Background

There is a rich body of work that addresses the design and implementation of network positioning systems [5–7, 12, 14, 15, 17, 18]. To be both practical and effective, a positioning system should be fully decentralized, incur scalable measurement overhead and provide reasonably accurate distance estimates. One approach to providing this scalability is to embed network distances in a low-dimensional coordinate space [5–7, 12, 14, 15, 18].

Among coordinate-based systems, the Vivaldi network positioning system [6] is the most widely deployed. It embeds network locations into an N-dimensional Euclidean space and fully decentralizes the computation of network locations. Each node maintains its own positions, periodically exchanges positions with other hosts and measures RTT latencies to them. The node then recomputes its position by simulating a force from a spring corresponding to the error between the Euclidean distance and measured distance. The authors evaluate the accuracy of the approach using PlanetLab nodes and King-based network distances between 1740 DNS servers and found that its error is competitive with

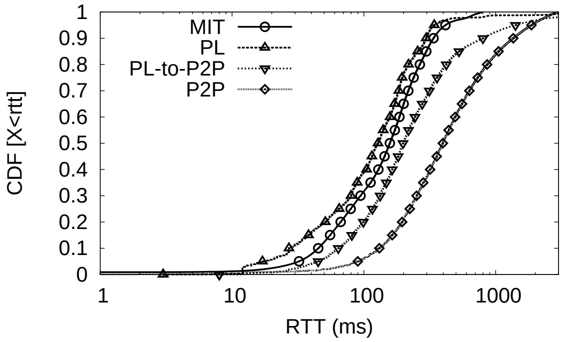

Figure 1: CDFs of latencies from different measurement platforms (semilog scale). Our measurement study exclusively between peers in Vuze (labeled P2P) exhibits double the median latency "in the wild" (labeled PL-to-P2P).

GNP [14]. In a follow-up study, Ledlie et al. [11] showed that the accuracy of this system was much lower "in the wild" as measured from PlanetLab nodes participating in the Vuze Vivaldi implementation. They proposed and implemented several features that improve accuracy in this environment, finding that accuracy improves by 43%. In fact, given the large scope of these systems (potentially thousands of networks), the constant evolution of the Internet and the time- and space-varying properties of latencies, it is a testament to the skillful design of these systems that they have achieved such relatively high levels of accuracy.

Despite the success of these systems, recent studies have called into question the usefulness of network coordinates [23]. For example, Wong et al. [22] note that embedding errors from network coordinates always leads to suboptimal peer selection and instead propose Meridian, a structured approach to direct measurement. To ensure scalability, Meridian organizes nodes into an overlay consisting of "rings" of nodes that locate nearby peers in log(N) steps, where N is the number of nodes participating in the system. Using a simulation-based evaluation with King-based latencies [10] between 2500 DNS servers and a deployment in PlanetLab with 166 nodes, the authors show that accuracy is significantly better than approaches that use virtual coordinates.

Using these the Vivaldi and Meridian approaches, this paper presents the first evaluation of network positioning performance from the perspective of tens of thousands of users in a real P2P setting.

3 Methodology

The results presented in this paper are based on measurements collected from more than 40,000 users broadly distributed worldwide. Because these peers are often located behind middleboxes at the edges of the network, they allow us to measure portions of the Internet not visible when using traditional measurement techniques [3].

For the purpose of finding low-latency peers in a DHT, Vuze concurrently maintains Vivaldi network coordinates using the original technique [6] and the one proposed by Ledlie et al. [11]. In addition to collecting these positions, our software performs ping measurements between connected peers, allowing us to compare each technique's distance estimate with ground truth. Finally, our software issues traceroute probes to collect topological information.

The advantage of our measurement approach is that it records measurements from the environment where network coordinates are used and at scale. As we demonstrate in Fig. 1, the distribution of latencies from our measurement platform (P2P) is much larger than those from MIT King [6] and PlanetLab (PL). In fact, the median latency in our dataset is twice as large as reported by Ledlie et al. [11], which used PlanetLab nodes to probe Vuze P2P users (PL-to-P2P).

The dataset used in this paper consists of over 1.54 billion Vivaldi samples, collected during the period of June 10 to June 25, 2008. After removing measurements that do not contain complete information (i.e., due to lack of ping response or uninitialized Vivaldi positions), we are left with over 1.41 billion Vivaldi samples from 43,674 IP addresses.

The ping measurements are further used to create a matrix for evaluating Meridian and to characterize the latency space as seen by end hosts. To ensure dense matrices for this analysis, we begin by bucketing our measurements into source and destination routable BGP prefixes (according to [19]), using the minimum observed RTT for each matrix element.[1] Because this approach still yields a sparse 6898x66825 matrix, we use the square submatrix and iteratively remove rows and columns that contain the largest number of empty elements until a sufficiently dense submatrix remains. Finally, the empty elements are filled with the median value for a given row to preserve that particular statistical property. We found that different approaches to filling empty matrix elements did not significantly affect the results when the matrix is nearly full. We use this process to generate a 495x495 matrix that is approximately 95% full. The rows represent ISPs in North America, Europe, Asia (including the Middle East), South America and Oceania.

Finally, we note that latency measurements are performed using the client's operating system ping command to prevent application-level latencies from affecting the measurements. We also note that while it

[1] We use the minimum to reduce the effect of latency variance on our analysis; however, as we show the results from our study still differ significantly from previous work.

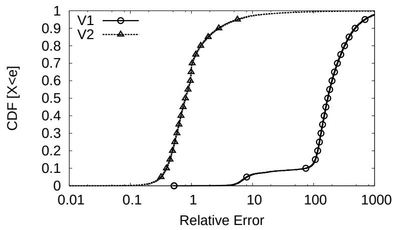

Figure 2: Absolute value of relative errors between estimated and measured latencies. Vivaldi V1 and V2 exhibit large median errors (1700% and 80%).

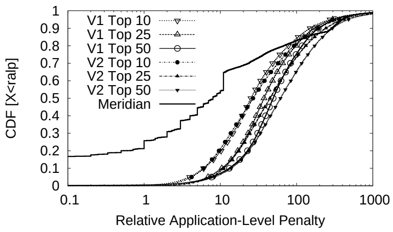

Figure 3: Relative application-level penalty for using network positioning. The vast majority of values are greater than one and the median values indicate order-of-magnitude loss in performance.

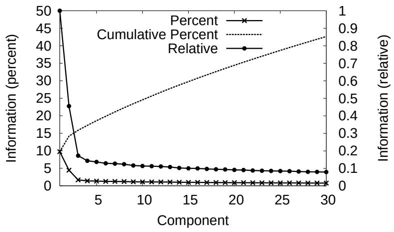

Figure 4: Plot indicating portion of variance captured by each principal component. The first few components capture only a small portion of the total variance.

is possible for client P2P traffic to interfere with the latency measurements (e.g., queuing delays due to large packages arriving at the user's router), this is precisely the kind of data that *any* network positioning system must account for in real deployments.

4 Performance from End Systems

In this section, we evaluate the accuracy of network positioning systems and their impact on the performance of an example application that uses them.

4.1 Accuracy

We begin our analysis by evaluating the accuracy the Vuze Vivaldi implementations in terms of estimated latency. Meridian is omitted here because it does not provide quantitative latency estimates. We focus on accuracy in terms of relative errors determined as follows: we first calculate the absolute value of the relative error between Vivaldi's estimated latency and the ping latency for each sample, then find the average of these errors for each client running our software. In Fig. 2, we plot a CDF of these values; each point represents the average relative error for a particular client. For the original implementation (labeled V1), the median relative error for each node is approximately 1700%, whereas the same for Ledlie et al.'s implementation (labeled V2) is 79.7% – both significantly higher than the 26% median relative error reported in studies based on PlanetLab nodes [11].

4.2 Impact on Applications

Relative error in latencies generally matter only if they negatively affect application performance. In the case of DHT performance, a positioning system need only guarantee that nodes that are closer to the target have smaller estimated distances than those that are farther away. One way to measure the extent to which this is true is the relative application-level penalty (RALP) metric initially proposed by Pietzuch et al. [16]. This metric measures the latency penalty incurred by applications using network positioning to select the closest N peers, compared to optimal selection.

To calculate RALP, we create an ordered set of node latencies according to ping measurements, P, and a set of nodes selected by the Vivaldi and Meridian systems, V. Then we find the average RALP for each measurement node using the following equation, where n is the number of nodes being measured and i is the index in the ordered sets:

$$1/n \cdot \sum_{i=1}^{n} (v_i - p_i)/p_i$$

Fig. 3 shows a CDF of the average RALP values for each measurement node when comparing the Meridian-selected node and the 10, 25 and 50 Vivaldi-selected nodes ordered by estimated latency. Note that the vast majority of RALP values is greater than 1, indicating that errors in the network positioning system lead to significant loss in performance for the DHT that uses it. For example, the median RALP for Vivaldi V2 when assessing the closest 10 nodes is 26.9, meaning that for half the peers in our study, the average latency to Vivaldi-driven peers is about 27 times worse than optimal. The situation worsens when evaluating the nearest 50 nodes – in this case, the median value for the average latency to those Vivaldi-recommended nodes is over 61 times worse than optimal. For Meridian, we simulate a network using our 495x495 latency matrix, adopting the same settings as the authors scaled to the scale of the network. Each value in the CDF is RALP, but for one selection. The results show that while more than 15% of the decisions are 100% accurate, the median error is 800%.

Based on the empirical results from our study, network coordinates not only exhibit large errors in predictions, but those errors significantly impact application performance in the Internet at large. In the next section, we explore why this is the case.

5 Sources of Error

Many authors have pointed out issues that impair accuracy in network positioning systems, including churn, coordinate drift, corruption, latency variance and intrinsic errors. While solutions have been proposed to address the first three problems [11], this section focuses on variance and intrinsic errors in latency prediction, as they represent fundamental challenges to every approach to network positioning.

5.1 Network Embedding

A popular approach to determining the intrinsic dimensionality of a system is to use principal component analysis (PCA). In the context of network latency, one constructs a matrix representing all-to-all latencies, then uses PCA to determine whether a small number of linear combinations of matrix elements represents most of its variance [18]. If the vast majority of the variance is modeled by a few principal components, then the network may be captured by a space containing a small number of dimensions. In various work, authors use this type of analysis to select 2, 4 or 7 dimensions [6, 11, 18].

To evaluate the intrinsic dimensionality of the Internet from the perspective of tens of thousands of end-users worldwide, we perform PCA on our latency matrix described in Sec. 3. In Fig. 4 we present a scree plot of the relative variance captured by each of the first 30 components, in descending order of the amount of variance they capture. The figure contains curves for *i*) the percent of the *total* variance captured by each component (**Percent**, left y-axis), *ii*) the *relative* variance captured by each component normalized by the value for the first component (**Relative**, right y-axis) and *iii*) the *cumulative* variance captured by all components with rank less than or equal to x (**Cumulative Percent**, left y-axis).

Traditionally, one uses the first two curves to identify the inherent dimensionality of the space by locating the "knee" in the curve. While the knee appears to be around the 4th or 5th component, these components capture only a small amount (18%) of the variance. Although the values quickly diminish for other components, the curve exhibits a long tail. For instance, 11 components are required to capture 25% of the variance and at least 41 components are required to capture 50% of the variance.

Previous work in PlanetLab has shown much higher variance captured by small numbers of coordinates, which can be explained by the platform's relatively small number of nodes located near the core. To hint at the effect of evaluating a smaller number of networks, we further reduced our matrix to 266x266 routable prefixes (99% full). After running PCA on this matrix, the amount of variance captured by the first few components *nearly doubles*. This suggests that there is more complex structure in the latency space seen from deployed P2P systems than is visible to limited deployments. We posit that this additional complexity is one of the primary reasons why network coordinates yield such large errors at scale, even with the improvements proposed by Ledlie et al [11].

5.2 Triangle Inequalities

Finally, we address the issue of triangle inequality violations (TIVs) in the Internet delay space caused by the network structure and routing policies. Wang et al. [21] demonstrate that high rates of TIVs (in their study, 12% of the triangles) significantly reduce the accuracy of network positioning systems. In other datasets, the occurrence of TIVs was relatively infrequent and thus many coordinate systems attempt to improve accuracy by filtering out TIV cases.

We performed a TIV analysis on our dataset and found that over 29% of the triangles had TIVs (affecting over 84% of the source/destination pairs) — this ranges from *over 4 times to an order of magnitude more TIVs* than reported in an analysis of datasets from Tang and Crovella [18] and it is significantly greater than that reported by Ledlie et al. [11]. While an important first step toward improving network positioning systems is to make them TIV-aware [21], it remains to be seen whether this approach can yield sufficient coverage and performance for client applications.

5.3 Variance and Last Mile Issues

It is well known that last-mile links often have poorer quality than the well provisioned links in transit networks; however, today's network positioning systems either ignore or naively account for this effect. To demonstrate the danger of ignoring this issue, Fig. 5 plots the portion of end-to-end latency along quartiles of the IP-level path between the measured hosts. We determine these statistics using 32M traceroutes and their per-hop latencies from nearly 68K IP addresses.

If the latency along a path were evenly distributed, the curves would center around $x = 0.25$. In fact, the first quartile (which is very likely to contain the first mile link) stands out from the rest, containing proportionately large fractions of the total end-to-end latency. For instance, when looking at the median values, the 1st quartile alone captures 80% of the end-to-end latency. The middle two quartiles, in contrast, each account for only 8%. Also note that the first quartile (and a significant fraction of the last quartile) has a large number of values close to and larger than 1. This

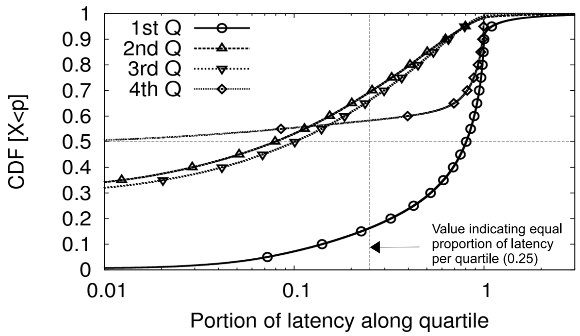

Figure 5: Plot indicating the portion of end-to-end latency contained in each quartile of the IP-level path between endpoints. The graph shows that the first quartile of the path contains the largest portion of latency most of the time, and the significant number of values greater than 1 indicate large variance in latencies in this portion of the path.

demonstrates the variance in latencies along these first and last miles, where measurements to individual hops along the path can yield latencies that are close to or larger than the total end-to-end latency (as measured by probes to the last hop). In fact, more than 10% of the 1st quartile samples have a ratio greater than 1.

While Vivaldi uses "height" to account for last-mile links [6] this analysis clearly shows that a single parameter is insufficient due to the large and variable latencies. The data instead suggests an approach that identifies links with high variance and treats them separately from the rest of the path.

6 Conclusion

In certain environments, e.g., those with low churn, low variance in latencies and small numbers of nodes, network positioning has shown to be extremely effective at accurately predicting network distances. However, we have shown that this is not the case when the systems are brought to much larger scale and run in residential networks worldwide. Large-scale network services (e.g., Akamai [2]) have addressed this by using topology information gathered from their networks. We believe that accounting for network topologies [13], while remaining fully decentralized, is an important next step in the evolution of distance estimation systems – an approach that we are currently investigating.

References

[1] ADLER, M., KUMARY, R., ROSSZ, K., RUBENSTEIN, D., SUEL, T., AND YAOK, D. D. Optimal peer selection for P2P downloading and streaming. In *Proc. of IEEE INFOCOM* (2005).

[2] AKAMAI. Akamai CDN. `http://www.akamai.com`.

[3] CASADO, M., GARFINKEL, T., CUI, W., PAXSON, V., AND SAVAGE, S. Opportunistic measurement: Extracting insight from spurious traffic. In *Proc. of HotNets* (November 2005).

[4] CHOFFNES, D. R., AND BUSTAMANTE, F. E. Taming the torrent: A practical approach to reducing cross-ISP traffic in P2P systems. In *Proc. of ACM SIGCOMM* (2008).

[5] COSTA, M., CASTRO, M., ROWSTRON, A., AND KEY, P. PIC: Practical internet coordinates for distance estimation. In *Proc. of the ICDCS* (2004).

[6] DABEK, COX, KAASHOEK, AND MORRIS, R. Vivaldi: A decentralized network coordinate system. In *Proc. of ACM SIGCOMM* (2004).

[7] FRANCIS, P., JAMIN, S., JIN, C., JIN, Y., RAZ, D., SHAVITT, Y., AND ZHANG, L. IDMaps: A global Internet host distance estimation service. *IEEE/ACM Transactions on Networking 9*, 5 (October 2001).

[8] FREEDMAN, M. J., FREUDENTHAL, E., AND MAZIÈRES, D. Democratizing content publication with coral. In *Proc. of USENIX NSDI* (2004).

[9] GUMMADI, K., GUMMADI, R., GRIBBLE, S., RATNASAMY, S., SHENKER, S., AND STOICA, I. The impact of DHT routing geometry on resilience and proximity. In *Proc. of ACM SIGCOMM* (2003).

[10] GUMMADI, K. P., SAROIU, S., AND GRIBBLE, S. D. King: Estimating latency between arbitrary Internet end hosts. In *Proceedings of Internet Measurement Workshop (IMW)* (November 2002).

[11] LEDLIE, J., GARDNER, P., AND SELTZER, M. Network coordinates in the wild. In *Proc. of USENIX NSDI* (2007).

[12] LIM, H., HOU, J. C., AND CHOI, C.-H. Constructing internet coordinate system based on delay measurement. In *Proc. of the Internet Measurement Conference (IMC)* (2003).

[13] MADHYASTHA, H. V., ANDERSON, T., KRISHNAMURTHY, A., SPRING, N., AND VENKATARAMANI, A. A structural approach to latency prediction. In *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement* (New York, NY, USA, 2006), ACM, pp. 99–104.

[14] NG, T., AND ZHANG, H. Predicting Internet network distance with coordinates-based approaches. In *Proc. of IEEE INFOCOM* (2002).

[15] PIAS, M., CROWCROFT, J., WILBUR, S. R., HARRIS, T., AND BHATTI, S. N. Lighthouses for scalable distributed location. In *Proc. of IPTPS* (2003).

[16] PIETZUCH, P., LEDLIE, J., AND SELTZER, M. Supporting network coordinates on PlanetLab. In *Proc. of WORLDS* (2005).

[17] SU, A.-J., CHOFFNES, D., BUSTAMANTE, F. E., AND KUZMANOVIC, A. Relative network positioning via CDN redirections. In *Proc. of the ICDCS* (2008).

[18] TANG, L., AND CROVELLA, M. Virtual landmarks for the internet. In *Proc. of IMC* (2003).

[19] TEAM CYMRU. The Team Cymru IP to ASN lookup page. `http://www.cymru.com/BGP/asnlookup.html`.

[20] VUZE, INC. Vuze, January 2009. `http://www.vuze.com`.

[21] WANG, G., ZHANG, B., AND NG, T. S. E. Towards network triangle inequality violation aware distributed systems. In *Proc. of IMC* (2007).

[22] WONG, B., SLIVKINS, A., AND SIRER, E. Meridian: A lightweight network location service without virtual coordinates. In *Proc. of ACM SIGCOMM* (2005).

[23] ZHANG, R., TANG, C., HU, Y. C., FAHMY, S., AND LIN, X. Impact of the inaccuracy of distance prediction algorithms on Internet applications - an analytical and comparative study. In *Proc. of IEEE INFOCOM* (2006).