The End-to-End Effects of Internet Path Selection

Stefan Savage, Andy Collins, Eric Hoffman
John Snell, and Thomas Anderson
Department of Computer Science and Engineering
University of Washington, Seattle

Abstract

The path taken by a packet traveling across the Internet depends on a large number of factors, including routing protocols and per-network routing policies. The impact of these factors on the end-to-end performance experienced by users is poorly understood. In this paper, we conduct a measurement-based study comparing the performance seen using the "default" path taken in the Internet with the potential performance available using some alternate path. Our study uses five distinct datasets containing measurements of "path quality", such as round-trip time, loss rate, and bandwidth, taken between pairs of geographically diverse Internet hosts. We construct the set of potential alternate paths by composing these measurements to form new synthetic paths. We find that in 30-80% of the cases, there is an alternate path with significantly superior quality. We argue that the overall result is robust and we explore two hypotheses for explaining it.

1 Introduction

In this paper we set out to explore a simple question: How "good" is Internet routing from a user's perspective, and why?

The impact of the Internet's routing protocols and policies on end-to-end performance is poorly understood. At any time there are many potential paths through the Internet connecting any two hosts. Some of these paths have higher bandwidth than others, some have lower propagation delay, and others see less congestion. These factors, which we call "path quality", ultimately limit the end-to-end performance achievable along any given path. As packets sent over a path are delayed or lost this directly reduces the throughput that a host can expect to obtain [MSM97].

Recent studies, such as Paxson's [Pax97a], have demonstrated that there is great diversity in the end-to-end performance observed on the Internet. However, it is currently unclear how much of this diversity should be attributed to differences in load, differences in capacity, or differences in the routing infrastructure. The focus of this paper is to explore this last possibility. Our goal is understand the degree to which end-to-end performance is being determined by the current state of Internet routing and to understand which mechanisms are responsible.

This work was funded by generous grants from NSF (CCR 94-53532), DARPA (F30602-98-1-0205), USENIX, the National Library of Medicine, Cisco, Fuji and Intel. Correspondence concerning this paper may be sent to savage@cs.washington.edu.

There are both technical and economic reasons to expect that Internet routing is non-optimal. Current wide-area routing protocols are primarily concerned with the exchange of connectivity information and do not incorporate measures of round-trip time or loss rate into their decisions. Per-network routing policies *can* be used to address inter-network performance concerns, but the combination of management complexity and the lack of economic motivation typically limit these policies to coarse grained intra-network load balancing [NAN]. Economic considerations can also limit routing options – some parts of the Internet refuse to carry traffic without a contractual agreement. However, it is unknown the extent to which any or all of these factors impact end-to-end performance as seen by the user.

To answer these questions we analyze five datasets containing large numbers of Internet path measurements taken between geographically diverse hosts. For the path connecting each pair of hosts, we collect several measures including round-trip time, loss rate, and, in one dataset, bandwidth. Using this framework, we compare each measured path to some potential alternates. These alternate paths are derived synthetically by composing the measurements from multiple connected paths. An optimal routing system would always choose the best available path between any two points on the Internet. The difference between the default and the synthetic alternate paths therefore is a rough measure of the efficiency of Internet routing. For 30 to 80 percent of the paths we examined, we find that there are alternates with significantly improved measures of quality.

Throughout this paper, we use the term "path" to refer to the complete set of hops traversed between two hosts, and the term "route" to refer to the data structures exchanged between routers to describe connectivity. While "route" is frequently used to represent both meanings, this can sometimes cause ambiguity. Similarly, we use the term "path selection" to describe the combined set of route selection decisions made at all the routers in a path.

In Section 2, we place our work in the context of related work concerning Internet routing behavior and end-to-end performance. In Section 3, we overview Internet routing protocols and policies. Section 4 describes the datasets we employed, our experimental and analytic methodologies, and some potential sources of error. We present our comparisons of Internet path quality in Section 5 and 6. Finally, we evaluate two hypotheses for explaining these results in Section 7, and we summarize our results in Section 8.

2 Related research

The relationship between path selection and end-to-end performance on the Internet has not been the subject of much study. While there is extensive analytic literature on routing protocols that guarantee a particular "Quality of Service" (e.g., [AGT98]), few papers address how path quality is affected by current Internet routing algorithms. Conversely there are many papers that measure the current Internet, but they do not consider how path quality is affected by routing decisions. However, these later papers are the inspiration for our work and are the most closely related.

The literature contains several studies that measure the behavior of Internet routing. Most of these focus on routing dynamics, that is, how routes change over time, and do not examine the issue of route selection. Chinoy's study of the NSFNET uses routing protocol traces to explore the frequency of changes in network connectivity [Chi93]. Using this data, Chinoy concludes that routing changes generally do not originate in the backbone, and that a small number of edge networks account for a disproportionate number of the total routing transitions. Labovitz et al. uses a similar methodology to examine pathologies in the observed behavior of the BGP routing protocol [LMJ97]. They find that the vast majority of routing updates are pathological and do not reflect real topological changes. Also, they show that periods of routing instability are correlated with periods of high traffic load and also exhibit strong periodicity. More recently, the same authors convincingly link some types of pathological routing behavior to the use of particular routing software implementations [LMJ99]. They further show that, discounting pathological behaviors, routing instability is well distributed across networks in the Internet and can not be easily attributed to a narrow class of networks. Paxson provides an end-to-end study of path dynamics by using the `traceroute` tool to identify the particular hops traversed between pairs of hosts [Pax96]. He finds that Internet paths are generally dominated by a single route, but that some networks do experience significant route fluctuation. Moreover, his data indicates that a large and increasing fraction of Internet paths follow different routes from source to destination than from destination to source.

Another area in which there is significant literature is the black-box study of Internet path characteristics. Bolot uses ICMP "echo" packets to examine the distributions of packet loss and round-trip times observed on a single trans-Atlantic path [Bol93]. A recent study by Paxson examines the characteristics of a larger set of paths using an automated analysis of TCP data transfers [Pax97a]. Paxson's results indicate that there is a wide variation in path characteristics such as round-time time, packet loss, and bandwidth. However, he also finds that the amount of available bandwidth tends to be stable for time periods up to several hours. Balakrishnan et al. also find significant temporal stability in bandwidth measurements collected from the IBM Olympic Web servers [BSSK97]. Further, they show that hosts which share portions of a path tend to obtain similar amounts of bandwidth.

There are a few papers that do touch explicitly on the interaction between Internet routing and path quality. McQuillan et al. describe a performance sensitive routing algorithm used in the early ARPANET [MRR80] and Khanna et al. [KZ88] discuss the behavior of this algorithm under varying degrees of load. Varadhan et al. present a simulation study of the effect of path changes on the performance of transport protocols [VEF98]. They show that small path changes during a TCP session can lead to significant reordering and a consequent reduction in performance. Finally, Francis et al. explore the possibility of using end-to-end measurements to construct maps of the minimum Internet propagation delay between hosts [FJP+99]. Their methodology is to predict the minimum propagation delay between a pair of hosts by triangulation using a series of pair-wise measurements. Their methodology, developed independently, is similar in principle to our approach of estimating the quality of an Internet path using synthetic alternate paths. As validation of our tool suite, we are able to independently generate their graphs.

3 Routing overview

Theoretically, if the Internet used "shortest" path routing, where paths are chosen to optimize some metric, there would be no room to find alternate paths with better performance. In reality, however, today's Internet routing policies and protocols are based on a number of factors that are only loosely correlated with performance.

The original ARPANET used a distributed adaptive routing algorithm based on measurements of queuing delay at each link. These measurements were propagated to all routers and packets were forwarded along the paths calculated to have the lowest delay [MRR80]. Early experience with this algorithm found that, under heavy load, routing oscillations made the system inefficient. Although more recent work has shown how to make performance-adaptive routing stable [Bre95], at the time the Internet resorted to a new metric of distance, "hop count," to be used during periods of high load [KZ88]. This metric correlates less well with performance than explicit measurements, but it tends to be more stable.

As the ARPANET evolved into an Internet connecting the networks of multiple agencies, the need for autonomous control emerged. Different agencies had their own backbones and wished to manage their internal routing differently from their connections with the Internet. This led to a two-level routing hierarchy that persists to this day. At the top-level, the Internet is partitioned into a relatively small number of autonomous systems (AS's). Routers within an AS route packets according to an interior gateway protocol (IGP); IGP's are used solely for selecting paths within an AS. Each AS is free to use its own metrics for selecting these internal routes. Although many small AS's (including the authors' home AS) still use raw hop counts to select internal routes, most larger AS's set internal metrics manually to distribute load and to avoid using links with excessive propagation delay [Fre98, NAN, Cor98].

Once a packet leaves an AS its routing is managed by a separate exterior gateway protocol (EGP) spoken in common by all other AS's. The first exterior gateway protocol was called EGP, and used a "hop count"-like metric based on the number of AS's in a path [Ros82]. The transition to a federation of regional networks and commercial backbones created new demands for independent control of routing policy. This resulted in the Border Gateway Protocol (BGP) used today [RL95].

Unlike the routing protocols described previously, BGP does not necessarily select routes by minimizing some global metric such as hop count or delay. Instead, the network administrators at each AS define a "routing policy" that dictates how routes are selected and advertised. This policy is implemented through a complex weighting scheme that allows an administrator to favor certain AS's to certain destinations, to encourage other AS's to favor certain exchange points, or to advertise a preference for being reached through one provider or another. However, in the absence of explicit policy rules, most BGP routers will select the routes with the shortest number of AS's in their advertisement.

Routing policies are driven by many concerns including, but probably not limited to: providing good end-to-end performance, addressing contractual obligations (e.g. "Acceptable Use Policies"), balancing load, minimizing cost, and incorporating local concerns about the quality of the routes provided by different providers. These policies not only control where a particular AS forwards traffic, but also affect how other AS's view the global topology. For instance, one AS may choose not to advertise certain routes, or may choose not to peer with another AS at all. Even between AS's that do exchange routes, a packet may not necessarily follow the “best” path. For example, a very common policy for large network service providers (NSPs) is “early-exit” routing [Fre98, Cor98]. In this scheme, traffic bound for another provider is routed to the nearest possible exchange point between them, whether or not this is the best path to the destination.

It should be clear that Internet routing is a complicated process and does not naturally lead to a performance-optimal selection of the path between two points. For a “good” path to be selected, the administrators of *every* AS on that path must have the incentive to maximize performance, must not have conflicting contractual or operational obligations, must possess the knowledge about what the best next hop is, must be able to express their knowledge in terms of policy, and must not be hindered by any other AS.

4 Methodology and measurements

One way of measuring the efficiency of Internet routing is to ask the question: “Is there an alternate path to our destination over which we would obtain better performance?” Unfortunately, while it is easy to directly measure the performance seen traversing the default path between two hosts, it is difficult to obtain the same metrics for alternate paths or even to discover what those alternate paths might be. The Internet does not have an effective mechanism that allows us to select the path taken by a packet; loose source routing, although an Internet standard, is disabled by many AS's because of security concerns. Nor does the Internet have a mechanism to reveal its complete internal connectivity: `traceroute`, a tool used in this study, only reveals internal links along the default path between the traced hosts.

Instead, we have opted to compare default paths to alternate paths we know must exist and for which we have reliable performance information – alternate paths using host-to-host paths as building blocks. We explain our methodology below, discuss some potential sources of bias, and then describe the actual datasets we have used to drive our analysis.

4.1 Methodology

The key observation behind our methodology is that different hosts have a different “view” of the network; they use different providers and have differing degrees of connectivity. Because end-to-end Internet paths are determined by the conjunction of a number of local routing policies, routing inefficiencies seen by one host are not necessarily seen by others. This allows us to compare the quality of the default path chosen by the Internet to a hypothetical path that routes around any inefficiency by traversing through a sequence of hosts.

To explore the quality of alternate paths, we collected three new large datasets of pair-wise host measurements taken over an extended period of time. We also used two other existing large datasets of pair-wise host measurements. We identify alternate paths by constructing a weighted graph in which each host is represented by a vertex and each path is represented by a corresponding edge. For all but one of the datasets, the weight of the edge is set according to the long term time average of the measurements (round-trip time, loss rate, or bandwidth) taken along that path over the length of the dataset. In one dataset, UW4-A, we repeatedly measure between all pairs of hosts at the same time, using these individual measurements as weights in the graph. In either case, for each pair of hosts, A and B, we remove the edge connecting them and perform a shortest-path computation between A and B using the remaining edges. The result is the best alternate path between A and B using other Internet paths as constituent “hops”. We repeat this experiment independently for each metric.

There are a number of potential biases in this approach, including both structural biases, those resulting from the choice of data, and statistical biases, those resulting from our analysis of the data.

While our ability to identify routing inefficiencies improves as the number of hosts increases, because we cannot measure potential routes using information about the internal Internet topology, our methodology will always yield a conservative estimate of the potential inefficiency in Internet routing. As just one example, many of our alternate paths traverse the same Internet links twice, on their way into and out of intermediate hosts; this cost would not be incurred by a real-life routing algorithm.

Another concern is about the representativeness of our data sources. In [Pax97b], Paxson presents an argument for the representativeness of two of the datasets we have used. For the other three datasets, we make no claim of representativeness; as with any trace-driven study, our results apply only to the hosts we measured at the times we measured them. Hosts in one dataset, UW1, were selected because they appeared in lists of public `traceroute` servers in North America; hosts in the other two datasets, UW3 and UW4, were selected because they were in North America and found to be `traceroute` servers by the Altavista search engine. All five datasets show qualitatively similar results, however, indicating that our results are probably not anomalous.

Another source of bias is that for all but one of our datasets, we do not measure all paths simultaneously, nor do we measure all hops on a single synthetic path simultaneously. We rely instead on long-term time averages of each metric of path quality (round-trip time, loss rate and bandwidth) to represent each default path, and combinations of these time averages to represent each synthetic alternate path. Consequently, each metric is influenced by samples taken at many different times of day and across many weeks. When we compare or combine two such statistics we are implicitly assuming that the measurements are all independent. This is clearly not true, as it is well established that many different parts of the Internet see higher load during weekday working hours and lower load during other times [TMW97]. In a few cases, this assumption of statistical independence is likely to yield a conservative estimate of routing inefficiency; for example, we assume loss rates are uncorrelated on each hop of an alternate path – correlated losses would only increase the relative benefit of the alternate path. In most other cases, the impact of the assumption of independence is uncertain.

In Section 6 we explore the impact of temporal dependence at multiple time scales. We first explore time-of-day variation and present data demonstrating that superior alternate paths are *more* prevalent during peak working hours. This leads us to surmise that dependence at this coarse time scale does not invalidate our results. There are also potential sources of dependence on shorter time scales, such as unfair buffer management or media access protocols. To address these concerns we have collected one dataset, UW4-A, in which all pairs are measured concurrently. We find superior alternate paths are somewhat *more* likely to be found when paths are measured simultaneously, although there is a large amount of variation in the performance of individual alternate paths at short time scales. This evidence suggests that at worst, the overall bias from our use of long-term averages is to *underestimate* the degree of routing inefficiency in the Internet.

An additional potential source of statistical bias arises from our use of the sample mean as a characteristic statistic. If the underlying distribution is skewed then the sample mean may be strongly affected. Despite this, we have chosen to use the mean because of the simplicity of its additive property (“the sum of the means is equal to the mean of the sums”) and the straightforward calculation of confidence intervals. Even assuming independence, the median

Dataset	Measurement method	Year collected	Duration	Location	Number of hosts	Number of measurements	Percent of paths covered
D2-NA	`traceroute`	1995	48 days	North America	22	14896	95
D2	`traceroute`	1995	48 days	World	33	35109	97
N2-NA	`tcpanaly`	1995	44 days	North America	20	7582	86
N2	`tcpanaly`	1995	44 days	World	31	18274	88
UW1	`traceroute`	1998	34 days	North America	36	54034	88
UW3	`traceroute`	1999	7 days	North America	39	94420	87
UW4-A	`traceroute`	1999	14 days	North America	15	216928	100
UW4-B	`traceroute`	1999	14 days	North America	15	9169	100

Table 1: Characteristics of the datasets used in this paper. "Percent of paths covered" represents the number of distinct paths measured divided by the number of potential paths that could have been measured (i.e., number of hosts * (number of hosts - 1)).

of the synthetic paths is substantially more expensive to compute than the mean. To do so requires that we convolve the samples of the edges being considered and extract the median of the resulting distribution. We have performed this analysis for a number of the graphs presented in this paper (we present one such graph in Section 6) and do not find any significant difference in the results.

4.2 Datasets

We conduct our analysis using several datasets whose basic characteristics are described in Table 1. In the remainder of this section we describe how the data was collected, concentrating on the new UW1, UW3, and UW4 datasets.[1] A more complete description of D2 and N2 can be found in [Pax96, Pax97a, Pax97b]. Note that D2 and N2 were collected three to four years earlier than the others and reflect a very different routing infrastructure.

All datasets used a centralized control host to generate requests to remote servers. In the UW datasets the remote servers were selected from publicly available `traceroute` servers, while D2 and N2 used a customized measurement daemon called `npd`. The control hosts issued requests to the servers at random intervals. In UW1 each `traceroute` server was chosen from a per-server uniform distribution with a mean of 15 minutes; the target of the `traceroute` was then chosen randomly from the list of servers. In UW3 and UW4-B, a random pair of hosts was selected for measurement using an exponential distribution with a mean of 9 and 150 seconds, respectively. In UW4-A, every server sent requests to every other server at the same time; these episodes were scheduled using an exponential distribution with a mean of 1000 seconds. Note that the 15 hosts in UW4-A and UW4-B are the same, and were selected at random from a pool of 35 hosts before the traces were started.

For the UW datasets, we empirically determined which hosts employed ICMP (i.e. `traceroute` reply) rate limiting, and filtered them from the datasets. Without such filtering, `traceroute` requests to rate limiting hosts would observe a higher loss rate than warranted. For UW1 we only removed such hosts from the pool of potential targets; instead, we use the round-trip measurements from `traceroutes` initiated in the opposite direction. For UW3 and UW4 we filtered all ICMP rate limiting hosts, to allow us to perform paired measurements on each path. For D2, identifying ICMP rate limiting hosts is no longer possible, so to correct for this bias we used the heuristic that only the first `traceroute` sample was counted against losses.[2] With the addition of this heuristic the distribution of loss rates in D2 became consistent with the other datasets. For the N2 dataset we only analyze the metric of bandwidth; since N2 measures round-trip time and loss rate observed within a TCP session, its measurements of those attributes are not unbiased samples. Finally, in D2 and in the UW datasets we removed paths for which there were fewer than 30 measurements so as to increase our confidence in the results.

These datasets have some potential biases and unique characteristics. First, in all datasets the control host was occasionally unable to contact the server it selected and this prevented a measurement from being made. In UW1, UW3, and UW4, measurements also failed if a request was not returned within 5 minutes. The consequence of this dependence between the control host and servers is to somewhat under-represent events correlated with host and server connectivity. In the context of our study, this causes us to overestimate the quality of intermittently or poorly connected paths. Second, the UW1 dataset was generated according to a uniform distribution and does not have the same theoretical protection against "anticipation" possessed by the datasets generated from the exponential distribution [Pax96]. This could potentially result in a reduction in the representativeness of the events in our data, but we can think of no anticipatory mechanism that would have a strong effect in this regard.

5 Results

Using the methodology described earlier, we evaluated the quality of alternate paths for the metrics of round-trip time, loss rate and bandwidth. Each graph presented in this section is a cumulative distribution function (CDF) across all pairs of hosts of the difference between the mean value for the metric in question and the mean value derived for the best alternate path for that metric; the alternate paths are selected according to a different metric in each graph. Values below zero (or below one for the relative graphs in Figures 2 and 5) are those for which the best alternate path was worse than the default path, while values above zero/one are those for which the best alternate path was superior. The distance from zero/one represents the magnitude of the difference. We have trimmed our graphs to eliminate visual scaling artifacts resulting from very long tails, so consequently some of our CDF's do not reach 100%. Uniformly, across all datasets and all metrics, we find that we are able to find good alternate paths between a significant fraction of the host pairs.

Figure 1 demonstrates this effect for the metric of round-trip time for the datasets of UW1, UW3, D2 and D2-NA. For 30 to 55 percent of the paths measured, there is an alternate path through

[1] A previous dataset, UW2, was removed from this paper due to uncorrectable experimental errors.

[2] Each `traceroute` invocation takes three consecutive samples of the round trip time to the end host; unless another `traceroute` was targeted to the same machine at the same time, the first sample to a rate-limiting host will be accurate, while the second and third samples are more likely to be dropped because they follow the first sample.

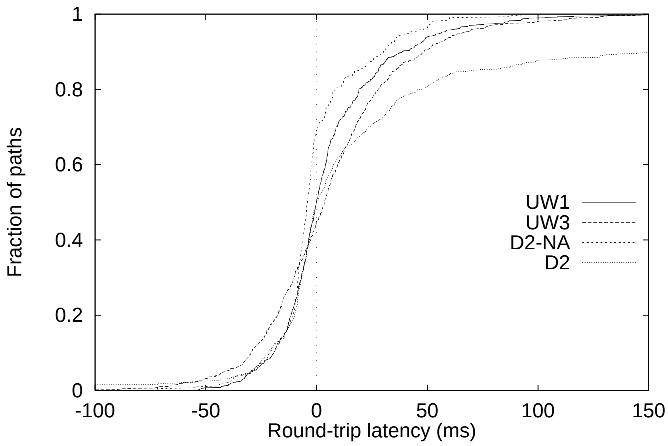

Figure 1: CDF of the difference between the mean round-trip time recorded on each path, and the best mean round-trip time derived for an alternate path.

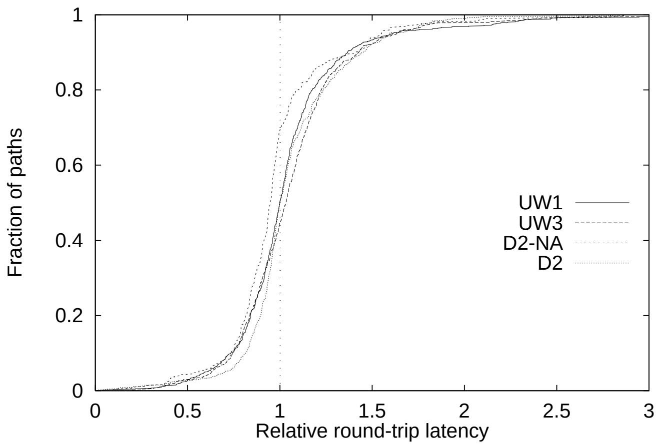

Figure 2: CDF of the *ratio* between the mean round-trip time recorded on each path, and the best mean round-trip time derived for an alternate path.

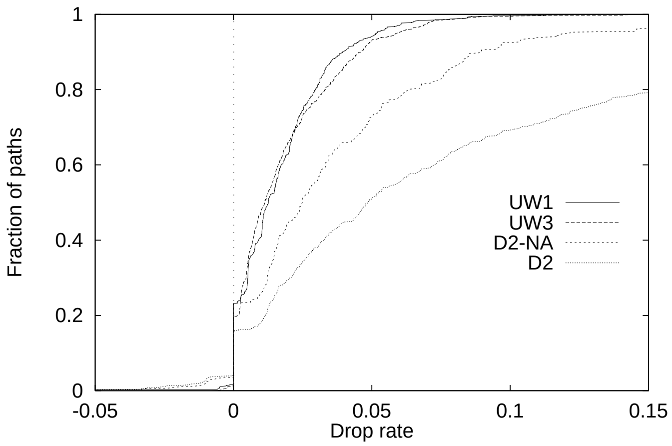

Figure 3: CDF of the difference between the mean loss rate recorded on each path, and the best mean loss rate derived for an alternate path.

one or more additional hosts resulting in a smaller round-trip time. For a smaller fraction, there was a significant improvement of 20 ms or more. Finally, when we take the *ratio* of the round-trip times for the default and best alternate path, shown in Figure 2, we find that for roughly 10 percent of the paths, the best alternate has 50 percent better latency. The imbalance between the D2 and D2-NA datasets in Figure 1 is due to the longer latencies for trans-oceanic transit; in Figure 2, the imbalance largely disappears.

A similar effect is demonstrated in Figure 3 for the metric of loss rate. Loss rates on synthetic alternate paths are formed by assuming that losses on the constituent "hops" are uncorrelated; an assumption of correlated losses would result in lower combined losses along alternate paths. Across all four datasets, we find that 75 to 85 percent of the paths have alternates with a lower loss rate. Again, the fraction of alternate paths that demonstrate substantial improvements in drop rate (5 percent or more) is smaller; only 5 to 50 percent of the paths fall in this category in the first three datasets. The vertical line at 0 percent represents pairs with no measured losses on either the default or alternate paths. Note that we did not collect enough samples to discriminate among low loss rates; we discuss confidence intervals for this graph in Section 6. For this same reason, normalizing the difference in the drop rate is uninteresting, as large numbers of alternate paths show enormous, or even infinite, relative improvements. As with round-trip time, most of the datasets track together, with D2 demonstrating substantially more improvement from alternate paths.

While the previous graphs suggest that there are alternate paths with better performance characteristics, they do not indicate the amount of available bandwidth on these paths. Although TCP performance is inversely related to background latency and drop rate, it is difficult to determine what the TCP throughput along an alternate path would have been from these measurements, because TCP exerts and reacts to load. Instead, we use the N2 datasets to attempt to answer this question, since they reflect the loss and round-trip times seen during actual TCP transfers. We construct alternate path bandwidth measurements by combining the round-trip times and loss rates observed along each default path in the N2 datasets. We compute the resulting TCP bandwidth according to the TCP model of Mathis et al. [MSM97]. We combine round-trip times via addition. However it is less clear how to compose loss rates, since we do not know how much of the observed loss was caused by the activity of the sending host and how much was due to background traffic. Therefore, we present the results using two different methods of combining loss rates. The first, which we label "optimistic", uses the maximum loss rate of any component of a synthetic path. This reflects the scenario that the sending TCP is completely responsible for the observed loss, and therefore the highest loss reflects the smallest bottleneck. The second, which we label "pessimistic", assumes that the loss rates on each component are independent and combines them according to the probability that a packet is lost on each underlying component of the synthetic path. This reflects a mode in which all of the measured packet losses are independent of the load exerted by the sending TCP. To be computationally tractable, we only consider alternate paths of length one hop for both the optimistic and pessimistic bandwidth metrics.

Using these procedures we compute the CDF of the difference between the bandwidth of the best alternate path and the actual measured bandwidth of the default path. Of course, since we do not have information about the capacity or load present on the links

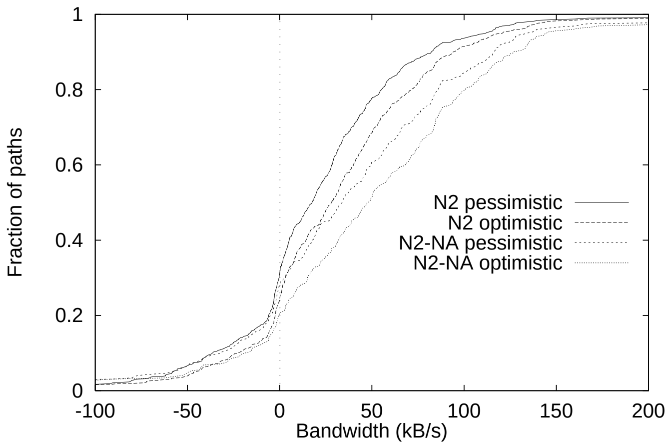

Figure 4: CDF of the difference between the mean bandwidth recorded on each path, and the best mean bandwidth derived for a one-hop alternate path. The lines labeled "optimistic" reflect alternate paths constructed using the maximum of the drop rates seen on the synthetic path, while the lines labeled "pessimistic" combine loss rates using the conditional probability assuming each rate is independent.

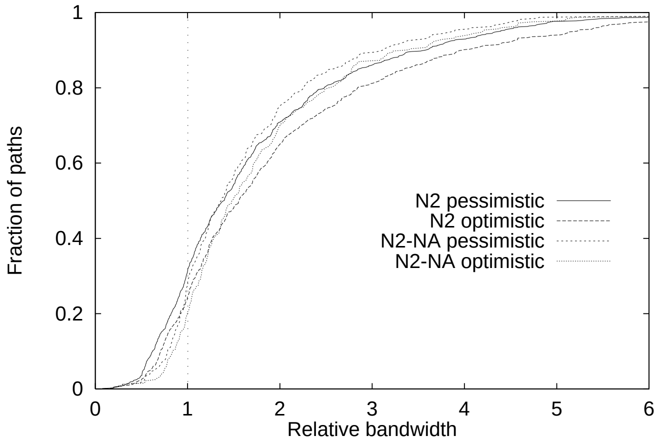

Figure 5: CDF of the *ratio* between the mean bandwidth recorded on each path, and the best mean bandwidth derived for a one-hop alternate path. The lines labeled "optimistic" and "pessimistic" reflect the same two cases as in Figure 4

of each path we cannot conclude that any difference would be significant for more than a single flow. The results in Figure 4 demonstrate that choosing alternate paths with respect to bandwidth shows the same pattern seen earlier; 70 to 80 percent of the paths have alternates with improved bandwidth. The optimistic and pessimistic curves provide a relatively tight bound for both datasets. Figure 5 shows the ratio of the computed alternate path bandwidth and the measured default path bandwidth. From this figure we can see that for at least 10% to 20% of the paths the potential bandwidth improvement is at least a factor of three. The difference between N2 and N2-NA in Figure 4 is due to the larger bandwidths available in North America; in Figure 5 the difference between the two datasets largely disappears.

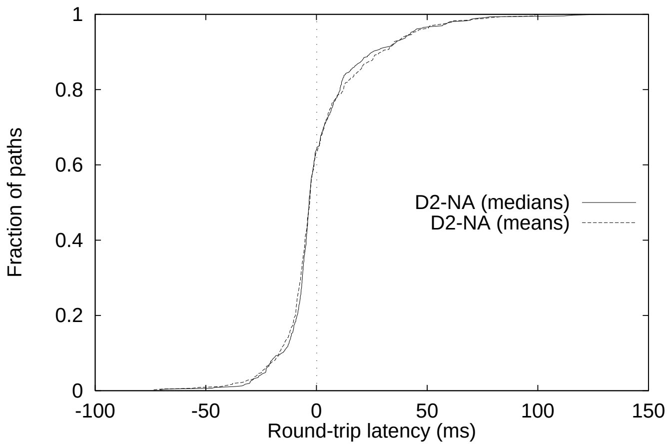

Figure 6: CDF of the difference between the mean round-trip recorded on each path, and the best mean round-trip time derived for any one-hop alternate path and the identical CDF using medians instead of means. This graph is for the D2-NA dataset.

6 Robustness

As discussed in Section 4, there are a number of biases in our methodology that might skew our results. In this section we evaluate the robustness our basic finding with respect to four of these factors: the use of the mean instead of the median, random variation among measurement samples, time-of-day dependence, and long-term averaging of path samples.

6.1 Use of the mean versus the median

The first issue we consider is the use of the mean instead of the median as our characteristic statistic. As discussed earlier, the mean may be affected if the underlying distribution is highly skewed. As a result, the median is usually considered a superior statistic. However, for our purposes, it is computationally expensive to compose distributions to yield median performance for synthetic alternate paths. We combine medians by convolving the distributions of the round-trip times in each path, and using the median of the resulting distribution. In Figure 6 we illustrate the difference between using the mean and the median when determining the alternate path improvement for round-trip time for one of the datasets. To keep the computational costs reasonable we limit the length of alternate paths for both means and medians to one hop. It is clear that for this metric on this dataset, the difference is negligible. We have sampled other datasets and metrics, and they all showed similar results.

6.2 Variation in the datasets

The second issue we consider is that of variation. All of our measurements demonstrate large ranges and consequently, it is possible that the difference between the means can be attributed largely to random variation in the data. Some of the potential sources of variation include upgrades to the network infrastructure during the traces, path changes (for instance due to routing policy changes or due to route flaps as in [LMJ97]), and congestion.

Following the procedure outlined in [Jai91], we compute the confidence interval for a single path as: $\bar{a} - \bar{b} \pm t_{[.975;v]}s$, where $\bar{a}$ and $\bar{b}$ represent the sample means for the path, $t_{[.975;v]}$ is the $(1 - \alpha/2)$-quantile of the t variate with v degrees of freedom, and

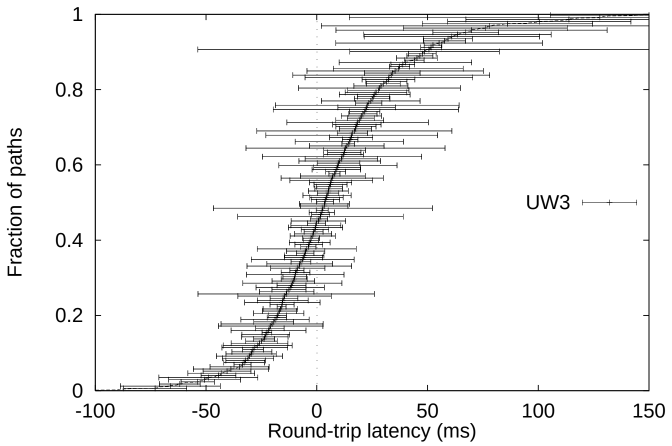

Figure 7: CDF of the difference between the mean round-trip time recorded on each path, and the best mean round-trip time derived for an alternate path. The 95% confidence interval is plotted as error bars for every eighth point on the y-axis. This graph is for the UW3 dataset.

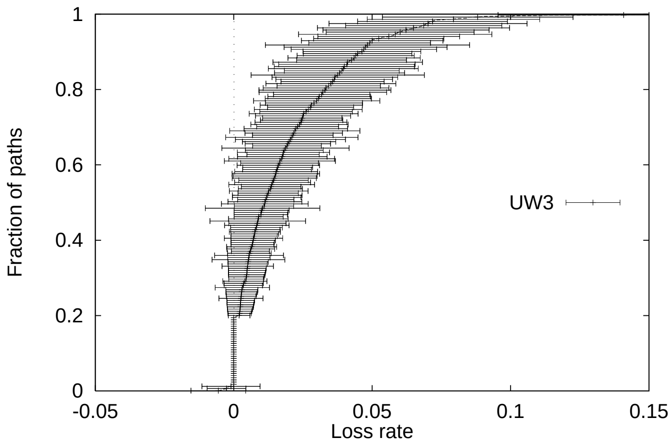

Figure 8: CDF of the difference between the mean loss rate recorded on each path, and the best mean loss rate derived for an alternate path. The 95% confidence interval is plotted as error bars for every eighth point on the y-axis. This graph is for the UW3 dataset.

s is the standard deviation of the mean difference. Note that this is computed independently for each point on the CDF.

In Figure 7 we plot the resulting 95% confidence intervals for the mean difference in round-trip time for the UW3 dataset; for readability, we include the intervals for every eighth path along the y-axis. In Figure 8 we do the same for loss rate. For round-trip time we see that although some paths have high variability, most paths have relatively tight error bounds. The same graph for loss rate shows larger variability; this is to be expected because each sample measurement of loss rate has a binary value, and consequently the standard deviation is quite large.

The confidence intervals show the same pattern across other datasets. In Tables 2 and 3 we show, for each dataset, the percent of the paths for which the best alternate path is better, worse, or indeterminate compared to the default path at a 95% confidence interval for round-trip time and loss rate. This is typically described as a *t-test* [Jai91]. Roughly speaking, the percentage of paths for which a better alternate path can be found at the 95% confidence level represents those paths whose improvement cannot be well explained simply by variation. This is a conservative measure of the effect of variation, because random variation could equally have been responsible for "hiding" alternate paths that were in fact better. While there is significant random variation, variation is not sufficient to explain the the difference between alternate and default paths.

Alternate is	UW1	UW3	D2-NA	D2
Better	28%	30%	20%	32%
Indeterminate	41%	41%	32%	37%
Worse	31%	29%	48%	31%

Table 2: Percentage of paths for which the difference in the mean round-trip time between the best alternate path and the default path is greater than zero, less than zero, or crosses zero at the 95% confidence level.

Alternate is	UW1	UW3	D2-NA	D2
Better	33%	46%	21%	41%
Indeterminate	42%	36%	57%	41%
Zero	21%	18%	19%	11%
Worse	4%	0%	3%	7%

Table 3: Percentage of paths for which the difference in the mean loss rate between the best alternate path and the default path is greater than zero, less than zero, is zero, or crosses zero at the 95% confidence level.

6.3 Time of day effects

Another concern is that we have averaged our data across large periods of time, and the quality of a path could vary significantly over the measurement period. For example, the Internet as a whole is likely to be more congested during peak working hours and less congested at night or on weekends. To investigate this concern, we have divided our data into weekday and weekend, and further divided weekday data into six hour time periods. In Figures 9 and 10, we show how this breakdown impacts the difference in the mean round-trip time and loss rate. For legibility we have only graphed data for the UW3 dataset, but the effects are similar for other datasets. Note that Figure 10 is not directly comparable to Figure 3 because dividing the dataset reduces the number of samples per path; this reduces our ability to discriminate the difference between default and alternate paths at low loss rates. This granularity effect is represented in the graph by the horizontal line that joins each curve to the vertical axis at 0% loss rate; it also reduces the tail where the default path outperforms the best alternate path.

The first thing we notice is that the overall effect occurs regardless of the time of day. It is also evident that time of day does impact the the magnitude of the difference. It is interesting to note that alternate paths seem to do better during times known to have heavier load. For both metrics, the greatest benefit is seen between the hours of 6am and 12pm PST, while the least benefit is seen during the weekend and between 12am and 6am PST. We hypothesize that during hours of low load there is little congestion or routing instability and there is less variance between paths. During periods of high load, we expect that routing instability and congestion are likely offering more "opportunities" for optimization.

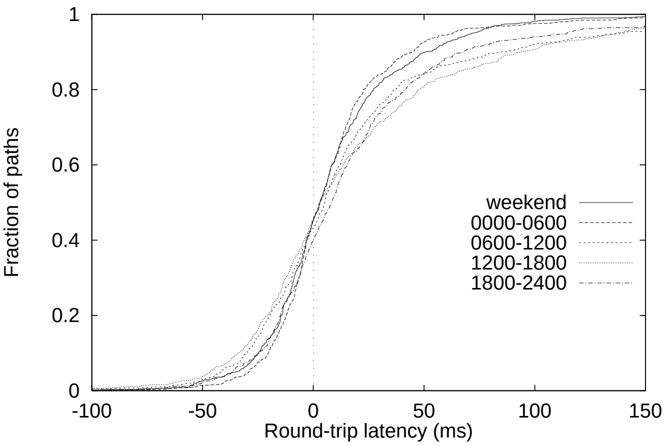

Figure 9: CDF of the difference between the mean round-trip time recorded on each path, and the best mean round-trip time derived for an alternate path, broken down by time of day and weekend. This graph is for the UW3 dataset.

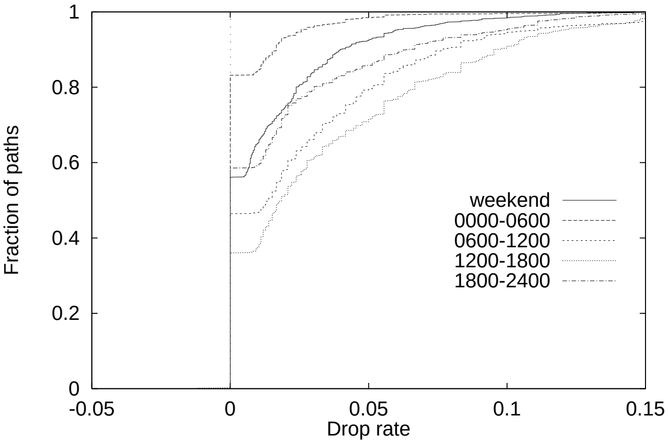

Figure 10: CDF of the difference between the mean loss rate recorded on each path, and the best mean loss rate derived for an alternate path, broken down by time of day and weekend. This graph is for the UW3 dataset.

6.4 Long-term averaging of data

For all of the data presented previously, the performance characteristics of each path were measured repeatedly over a relatively long timescale, and averaged together before any comparisons were made to potential alternate paths. In order to gauge the effect of this averaging, we took a new dataset, called UW4-A, for which we measured all paths concurrently. Out of a pool of 35 of the hosts used in UW3, we randomly selected 15 to use for this measurement. UW4-A consists of a series of randomly spaced "episodes," and each episode consists of a single `traceroute` measurement in each direction between each pair of the 15 hosts. Of course, each `traceroute` request takes a non-negligible amount of time to execute, because it measures the round-trip time to each intermediate router before reaching the target host. Therefore our measurements are "simultaneous" only within a several minute window. In analyzing UW4-A, we compute the best alternate path using only measurements taken from the same episode; we then calculate the

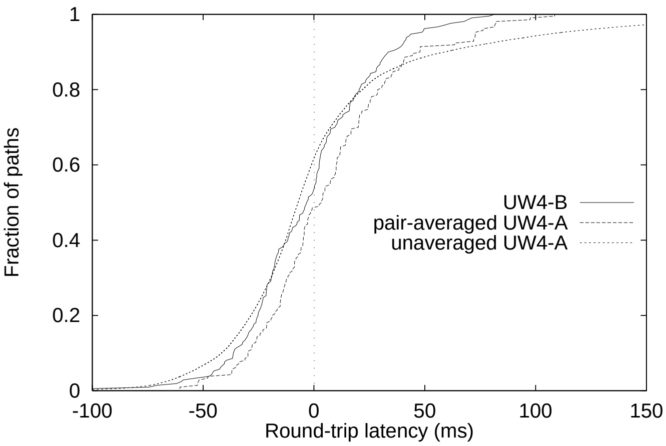

Figure 11: CDF of the difference between the mean round-trip time recorded on each path, and the best mean round-trip time derived for any alternate path (labeled "UW4-B"), plus the equivalent CDF using the mean difference for each pair of hosts between the default path and best alternate path when measurements are made "simultaneously" on all paths (labeled "pair-averaged UW4-A"), plus the equivalent CDF using the raw individual differences between the default path and the best alternate path (labeled "unaveraged UW4-A").

difference between the measurement of the default path and the best alternate path within the episode. To serve as a basis of comparison, during the period we collected UW4-A, we also made an independent set of long-term time average measurements between the same set of 15 hosts; we call this dataset UW4-B.

Figure 11 plots the resulting comparison between the simultaneous measurements in UW4-A and the time-averaged measurements in UW4-B. We plot the simultaneous measurements in two separate ways. One curve, labeled "pair-averaged," chooses the best alternate path for each pair of hosts for each episode, and averages the resulting difference across all measurements for the same pair. This is the curve that is most comparable to the time-averaged mean from UW4-B, in that each data point in the graph represents the average performance seen by a single pair of hosts. The result shows that we are slightly more likely to be able to find good alternate paths on a fine-grained timescale than on a long-term timescale.

We should note, however, that there is a huge amount of variability in the performance of the best alternate paths in UW4-A. For a given pair of hosts, not only are different alternate paths being selected as best in each episode, the difference between the best alternate path and the default path is highly variable. Many of the pairs of hosts have large swings from episode to episode, where the best alternate path is sometimes much worse and sometimes much better than the default path. To capture this variability, we also plotted in Figure 11 the CDF of all the individual measurements of the differences between the best alternate path and the default path. This curve, labeled "unaveraged", plots a data point on the CDF for every pair of hosts for every episode. The graph shows a much broader tail in both directions when the points are plotted individually.

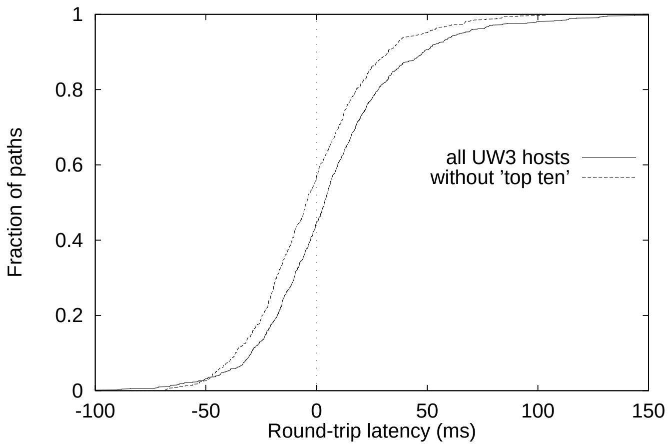

Figure 12: CDF of the difference between the mean round-trip time recorded on each path and the best mean round-trip time derived for an alternate path, and the equivalent CDF computed for the dataset after having removed the "top ten" hosts. This graph is for the UW3 dataset.

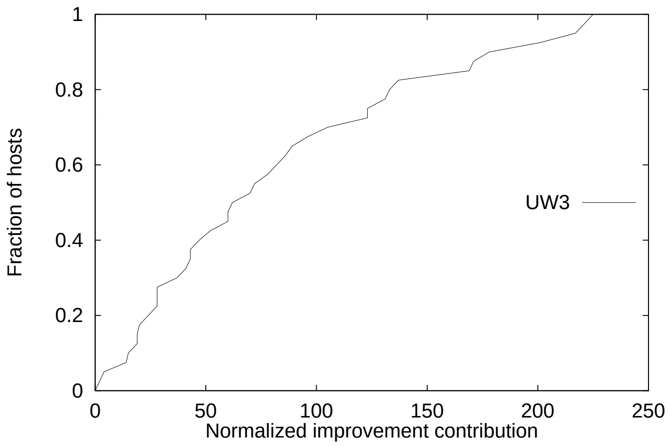

Figure 13: CDF of the number of better alternate paths in which a host appears as an intermediate node, weighted by the degree to which the alternate path is better

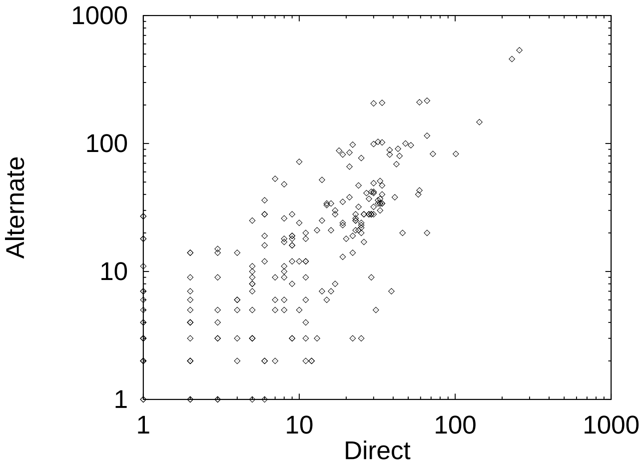

Figure 14: Scatterplot of AS's found in the UW1 dataset. The x-axis represents the number of default paths in which that AS appears, while the y-axis is the number of best alternate paths (for the metric of round-trip time) in which it appears.

7 Evaluation

In this section we evaluate two hypotheses for explaining the presence of superior alternate paths. The first hypothesis is that superior alternate paths are caused by avoiding parts of the Internet with particularly poor quality (e.g. congested exchange points) or by exploiting connectivity to parts of the Internet with exceptionally good quality (e.g. vBNS). The second hypothesis is that, more specifically, superior alternate paths result primarily from avoiding congestion, rather than by minimizing propagation delay. We discuss each of these theories below.

7.1 Host and AS popularity in alternate paths

To evaluate the degree to which the prevalence of superior alternate paths is a result of the behavior of a small part of the Internet or a more widespread phenomenon, we conducted several experiments that attempt to quantify the effect that an individual host or a single autonomous system (AS) can have on our results.

Our first experiment evaluates the effect of removing individual hosts from a dataset in terms of the shape of the alternate path CDF curve. If it were the case that only a handful of nodes were somehow causing the existence of most of the superior alternate paths, then we should be able to remove those hosts and see a dramatic shift of the CDF curve for the remainder of the dataset.

Figure 12 shows the effect of removing the ten hosts which have the greatest impact on the CDF curve. We use a simple greedy algorithm to select the hosts; at each step we remove the host whose removal shifts the CDF the farthest to the left. The curve labeled "all UW3 hosts", previously reported in Figure 1, shows the distribution of the absolute improvement in round-trip time between the best alternate path and the default path for the entire UW3 dataset, while the curve labeled "without 'top ten'" shows the distribution after removing the top ten hosts. From the figure, we see that the top ten hosts are not the source of a disproportionate number of the superior alternate paths, and we conclude that the prevalence of alternate paths with superior round-trip times cannot be attributed to a small number of hosts.

We next measure the number of times each host appears as an intermediate host in some superior alternate path (not necessarily the very best alternate), weighted by the degree to which the alternate path was better than the corresponding default path. Figure 13 shows the CDF of this "normalized improvement contribution" for each host in the UW3 dataset. We can see that the distribution lacks the heavy tail that would indicate the existence of a few hosts with abnormally large contributions, so again we cannot attribute the existence of superior alternate paths to a small number of hosts.

Finally, we consider the effect of autonomous systems in the center of the network, rather than individual end hosts in our datasets. For each AS that appeared in any trace in the dataset, we compute the number of default paths in which that AS appears and the number of best alternate paths in which it appears. Figure 14 shows a scatter plot relating these two quantities for the metric of round-trip time for the UW1 dataset; each point represents a single AS in the dataset. Since the graph does not show a significant number of AS's which are substantially more represented in either the original paths or the alternates, we conclude that the availability of alternate paths is not being unduly inflated by a small number of

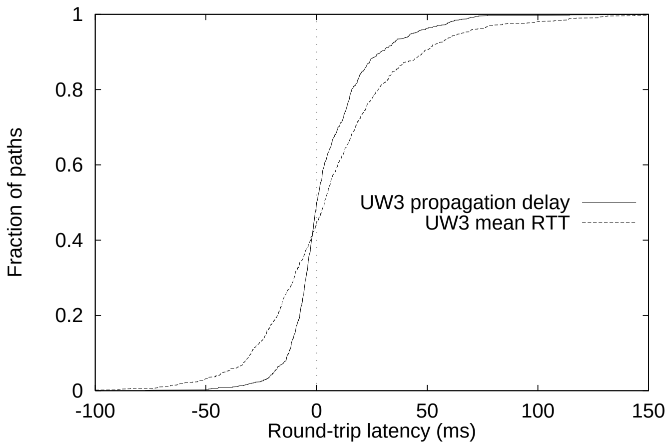

Figure 15: CDF of the difference between the propagation delay recorded for each path, and the best propagation delay derived for an alternate path, superimposed with the equivalent CDF for the mean round-trip time. This graph is for the UW3 dataset.

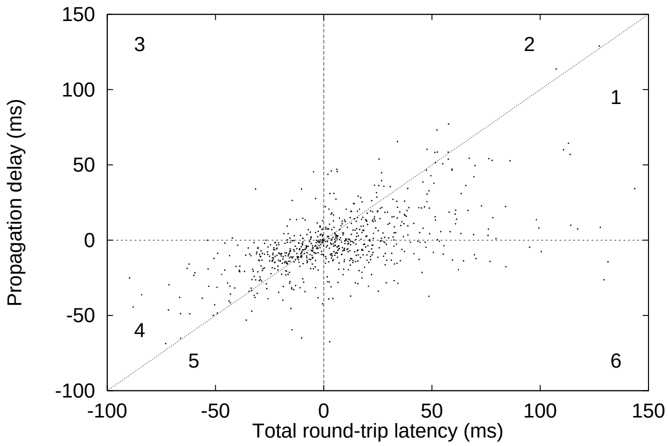

Figure 16: Scatterplot of the portion of the difference in mean round-trip time between the best alternate path and the default path due to the difference in physical propagation delay. Each point represents one pair of hosts. This graph is for the UW3 dataset.

either good or poor AS's.

7.2 Congestion vs propagation delay

In the preceding sections, we used mean round-trip time as our measure for network latency, with the goal of capturing the end-user performance of different paths. In this section we will consider the relationship between the two components of mean round-trip latency: propagation delay and queuing delay. Propagation delay includes all fixed costs along the path, primarily physical transmission latency, minimal store and forwarding delay, and processing overhead; queuing delay corresponds to the congestion-dependent costs. Although we cannot directly measure propagation delay, we can estimate it from our data by taking the tenth-percentile of the measured round-trip times. We chose to take the tenth percentile rather than the actual minimum observation to protect against noise in the case where the minimum resulted from a different route than the majority of the measurements.

If congestion is a major source of routing inefficiency and if avoiding congested links is a major reason for the existence of superior alternate paths, then two hypotheses should hold. We should find less inefficiency with respect to propagation delay than we saw with respect to mean latency, and we should find that most of the gains for alternate paths for mean latency to be due to reduced queuing delay rather than reduced propagation delay.

Figures 15 and 16 show that while both of these hypotheses are true to a limited degree, there is still a substantial degree of inefficiency in propagation delays and a substantial tendency for paths with superior mean round-trip latencies to also have superior underlying propagation delays. Figure 15 plots the CDF of the difference between the best alternate path and the default path using the metric of propagation delay, for the UW3 dataset. This figure was generated using the same methodology described in Section 4, but substituting propagation delay as the metric by which alternate paths are chosen and judged. A CDF for the same dataset with respect to mean round-trip latency is included from Figure 1 for comparison. From Figure 15 we can see that, although the magnitude of the differences is cut substantially when only propagation delay is considered, superior alternate paths still exist for 50% of the paths, and the differences are still significant for a considerable number of paths.

Figure 16 looks at the relative contributions of propagation and queuing delay for the default and best alternate paths, again for the UW3 dataset. For this figure, the alternate paths were selected with respect to the mean round-trip time. The difference in the round-trip time between the best alternate path and the default path was then separated into the portion due to the propagation delay versus the queuing delay. Each point in the scatter plot corresponds to a single path; the x axis measures the difference in mean round-trip latency between the default and best alternate paths, and the y axis measures just the difference in propagation delay. Points to the right of the y axis are paths where a superior alternate path exists, while for those on the left the default path is superior.

If it were the case that superior paths are superior entirely because they avoid congestion, then we would expect to see the points clustered around the x axis, with no positive correlation between total latency and propagation delay. Conversely, if congestion was equal along all paths and all of the improvements were due to propagation delay, we would see all the points clustered around the line $y = x$. What we see in the data is a mixing of these properties, with the points where the default path is superior falling primarily between the x axis and the line $y = x$, and the points where the alternate path is superior tending a little more towards the x axis. These effects vary somewhat between the different datasets.

The points are separated into six qualitative groups by the two axes and the line $y = x$. Each group is largely symmetric with its reflection about the origin, with the primary difference being whether it is the default or the alternate path that is superior. Points in groups 1 and 4 are what might be considered "typical" points, where the better path is superior both in queuing delay and propagation delay. In groups 2 and 5, the difference in propagation delay is *greater* in magnitude than the difference in mean latency, which indicates that the queuing delay is actually *worse* along the superior path, while in groups 3 and 6 the difference in propagation delay is *opposite* the difference in mean latency, indicating the the superior path has greater propagation delay, and therefore *much* smaller queuing delay. As might be expected, there are very few paths in group 3, indicating that most superior *default* paths have better propagation delay, while group 6 is much more populated, indicating that many superior alternate paths are in fact going out of their way to avoid congestion.

The conclusion from this data is that congestion and propagation delay both play significant roles in the observed inefficiencies; neither one can properly be said to be the single dominant factor.

8 Conclusion

In this paper, we have presented a methodology for finding and measuring the potential performance of alternate paths through the Internet. We have shown that for a large number of paths in the Internet there are alternate paths that exhibit superior quality as measured by round-trip, loss rate, and bandwidth. We have argued that this finding is a robust one, largely independent of the precise set of hosts measured, and applying to datasets taken across a several year period, at different times of the day, whether instantaneous or long-term time average performance is examined, and whether the minimum delay or the mean round-trip time is considered.

Acknowledgments

We would like to thank the following individuals for their contributions to this project. Beatrix Jones and Ronit Katz provided extensive advice on all things statistical. Vern Paxson supplied us with the D2 and N2 datasets and gave extensive feedback concerning their analysis as well as general advice on network measurement. Neal Cardwell, Scott Shenker, Amin Vahdat, David Wetherall, and the anonymous SIGCOMM reviewers all provided valuable feedback on earlier drafts of this paper. Finally, we would like to thank the maintainers of the public `traceroute` servers we used to conduct our measurements; without their good will this study would have been much more difficult.

References

[AGT98] George Apostolopoulos, Rouch Guerin, and Sanjay Kamatand Satish Tripathi. Quality of Service Routing: A Performance Perspective. In *Proceedings of the ACM SIGCOMM '98*, pages 17–28, Vancouver, BC, September 1998.

[Bol93] J-C. Bolot. End-to-end Packet Delay and Loss Behavior in the Internet. In *Proceedings of the ACM SIGCOMM '94*, pages 289–298, San Francisco, CA, September 1993.

[Bre95] Lee Michel Breslau. *Adaptive Source Routing of Real-Time Traffic in Integrated Services Networks*. PhD thesis, University of Southern California, Department of Computer Science, December 1995.

[BSSK97] Hari Balakrishnan, Srinivasan Seshan, Mark Stemm, and Randy H. Katz. Analyzing Stability in Wide-Area Network Performance. In *Proceedings of the 1997 ACM SIGMETRICS Conference*, Seattle, WA, June 1997.

[Chi93] Bilal Chinoy. Dynamics of Internet Routing Information. In *Proceedings of the ACM SIGCOMM '94*, pages 45–52, San Francisco, CA, September 1993.

[Cor98] Steve Corbato. Personal communication, 1998.

[FJP$^+$99] Paul Francis, Sugih Jamin, Vern Paxson, Lixia Zhang, Daniel Gryniewicz, and Yixin Jin. An Architecture for a Global Internet Host Distance Estimation Service. In *Proceedings of the IEEE INFOCOM '99*, New York, NY, March 1999.

[Fre98] Avi Freedman. Optimal External Route Selection: Tips and Techniques for ISPs. Tutorial at 14th North American Network Operators Group (NANOG) Meeting, November 1998.

[Jai91] Raj Jain. *The Art of Computer Systems Performance Analysis*. Wiley Professional Computing, 1991.

[KZ88] Atul Khanna and John Zinky. The Revised ARPANET Routing Metric. In *Proceedings of the ACM SIGCOMM '88*, pages 45–56, Palo Alto, CA, August 1988.

[LMJ97] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. Internet Routing Instability. In *Proceedings of the ACM SIGCOMM '97*, pages 115–126, Cannes, France, September 1997.

[LMJ99] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. Origins of Internet Routing Instability. In *Proceedings of the IEEE INFOCOM '99*, New York, NY, March 1999.

[MRR80] J.M. McQuillan, I. Richer, and E.C. Rosen. The New Routing Algorithm for the ARPANET. *IEEE Transactions on Communications*, 28(5):711–719, May 1980.

[MSM97] Matthew Mathis, Jeffrey Semke, and Jamshid Mahdavi. The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. *ACM Computer Communications Review*, 27(3):67–82, July 1997.

[NAN] North American Network Operators Group. *North American Network Operators Group (NANOG) mailing list*. `http://www.nanog.org`.

[Pax96] Vern Paxson. End-to-End Routing Behavior in the Internet. In *Proceedings of the ACM SIGCOMM '96*, pages 25–38, Stanford, CA, August 1996.

[Pax97a] Vern Paxson. End-to-end Internet Packet Dynamics. In *Proceedings of the ACM SIGCOMM '97*, pages 139–152, Cannes, France, September 1997.

[Pax97b] Vern Paxson. *Measurements and Analysis of End-to-End Internet Dynamics*. PhD thesis, University of California at Berkeley, Department of Electrical Engineering and Computer Science, April 1997.

[RL95] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). RFC-1771, 1995.

[Ros82] E. Rosen. Exterior Gateway Protocol (EGP). RFC-827, 1982.

[TMW97] Kevin Thompson, Gregory J Miller, and Rick Wilder. Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network Magazine*, 11(6):10–23, November 1997.

[VEF98] Kannan Varadhan, Deborah Estrin, and Sally Floyd. Impact of Network Dynamics on End-to-End Protocols: Case Studies in TCP and Reliable Multicast. Technical Report USC-CS-TR 98-672, University of Southern California, Information Sciences Institute, April 1998.