Quantifying the Causes of Path Inflation

Neil Spring Ratul Mahajan Thomas Anderson

{nspring,ratul,tom}@cs.washington.edu
Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350

ABSTRACT

Researchers have shown that the Internet exhibits path inflation – end-to-end paths can be significantly longer than necessary. We present a trace-driven study of 65 ISPs that characterizes the root causes of path inflation, namely topology and routing policy choices within an ISP, between pairs of ISPs, and across the global Internet. To do so, we develop and validate novel techniques to infer intra-domain and peering policies from end-to-end measurements. We provide the first measured characterization of ISP peering policies. In addition to "early-exit," we observe a significant degree of helpful non-early-exit, load-balancing, and other policies in use between peers. We find that traffic engineering (the explicit addition of policy constraints on top of topology constraints) is widespread in both intra- and inter-domain routing. However, intra-domain traffic engineering has minimal impact on path inflation, while peering policies and inter-domain routing lead to significant inflation. We argue that the underlying cause of inter-domain path inflation is the lack of BGP policy controls to provide convenient engineering of good paths across ISPs.

Categories and Subject Descriptors

C.2.1 [**Communication Networks**]: Architecture and Design—*topology*

1. INTRODUCTION

In this paper, we attempt to answer a simple question: why are Internet paths sometimes absurdly long? We see quantifying the causes of path inflation as a step toward the broader goal of understanding the factors that shape Internet routes. We start with the observation that in a well-provisioned and well-operated network, direct shortest paths are preferred. Intentionally longer paths result from the interaction of topology and policy at three layers – in the selection of paths within an ISP, of peering links to reach neighboring ISPs, and of the sequence of ISPs used to reach more distant destinations. Thus, characterizing the extent to which different factors inflate paths helps us to gain insight into the design of the network, routing protocols and ISP policies. For instance, we found that a large fraction of packets entering AT&T at San Francisco and destined for Sprint experienced significant path inflation. Closer investigation revealed that this was caused by routing policy intended to avoid a congested peering link.

Over the past few years, researchers have found significant path inflation in the Internet [1, 11, 26, 32]. They have postulated possible causes and, in some cases, ruled out others [32], but no prior work has completely explained the effect. This is challenging because most ISPs are unwilling to share accurate, up-to-date information about their topology or routing policy, considering such matters to be proprietary. Fortunately, advances in Internet topology mapping [7, 15, 27] and ISP policy inference [12, 29] have begun to make this information more widely available.

In this paper, we identify six possible causes of path inflation – topology and routing policy at all three layers mentioned above – and quantify the relative impact of each. We start by measuring the POP-level backbone topology of 65 diverse ISPs and their interconnections, by tracing from 42 vantage points to all globally-routed IP address prefixes. We then develop new methods to infer intra-domain routing policies and ISP peering policies (which determine how ISPs exchange traffic with their neighbors). Our key insight is that both paths taken and not taken reveal information about the routing policy. For intra-domain paths, we infer a set of edge weights that are consistent with the paths observed in the trace data. Between neighboring ISPs, we infer the use of early exit ("hot potato") and various forms of cooperative routing by focusing on whether the chosen paths depend on the ingress in the upstream ISP, the egress in the downstream ISP, or both. These results provide the first empirical data on ISP peering policies.

Using a trace-driven methodology, we analyze observed and hypothetical topologies and policies to isolate the impact of the various factors on path inflation. We find that i) intra-domain traffic engineering is commonplace but has only a minimal impact on path inflation; ii) contrary to popular belief, there is significant cooperation (helpful non-early-exit) between adjacent ISPs in the Internet, to avoid particularly poor routes or load-balance traffic across multiple peering links; iii) many paths that use early-exit are inflated compared to a hypothetical optimal exit policy; iv) topology-insensitive load balancing can cause significant path inflation. We also confirm earlier work [26, 32], showing that roughly half of all paths are inflated due to inter-domain routing; this inflation arises not from commercial constraints, but from using AS-path length as a routing metric. Because our study is trace-driven, our results are limited to our choice of vantage points and ISPs, as well as by a lack of visibility into provisioning and policy choices made below the IP layer. Further work will be needed to determine the extent to which our results generalize to other periods of time and other vantage points.

SIGCOMM'03, August 25–29, 2003, Karlsruhe, Germany.

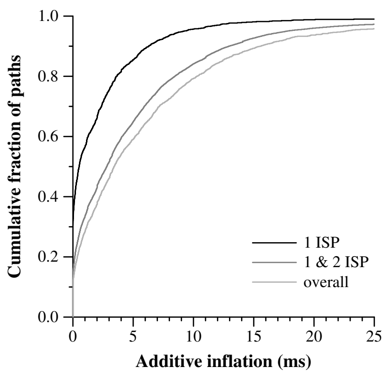

Figure 1: Path inflation observed in our dataset as a cumulative distribution. The x-axis represents the extra distance traveled through the network beyond a hypothetical direct link between end points. *1 ISP* paths stay within one ISPs network; *1 & 2 ISP* paths traverse a maximum of two ISPs; *overall* is the complete dataset.

Overall, our results show that particularly long paths are due neither to incompetent network engineering nor hypercompetition between ISPs. Instead, a main culprit is the design of BGP, the current inter-domain routing protocol, that makes it difficult to implement robust topology-sensitive routing decisions. For instance, there is no easy way for two adjacent ISPs to cooperate to implement optimal exit routing, should they desire to do so.

In Section 2, we categorize the root causes of path inflation into the six factors we study in the rest of the paper. Section 3 outlines our measurement and analysis methodology. Section 4 focuses on intra-domain topology and policy, Section 5 on the peering links and policy between neighboring ISPs, and Section 6 on the impact of inter-domain topology and policy on path inflation. We summarize our results in Section 7, discuss related work in Section 8, and conclude in Section 9.

2. THE CAUSES OF PATH INFLATION

While the direct optical fiber latency between San Francisco and Boston is 20 ms, the usual one-way network delay is much more. This is not a unique situation. Considering pairs of cities, Figure 1 illustrates the prevalence of path inflation in our data set relative to a hypothetical direct link.

Our goal is to analyze the factors that cause inflation to better understand the impact of network engineering and routing policy on path selection. Network engineering, the design of the topology of the network, can cause inflation if there are too few direct links in the network graph. Routing policy can cause inflation because existing shorter paths may not be selected to carry traffic. Path inflation provides a metric for study that allows relative comparisons between these very different factors.

We characterize the contribution of topology and policy to path inflation at each of the three component layers described below.

1. Intra-domain: The Internet is composed of thousands of inter-connected but independent administrative domains known as ISPs or autonomous systems (AS). "Intra-domain" refers to parts of the network that belong to a single ISP. Both intra-domain topology and policy can lead to path inflation; the combined impact of these two factors is depicted by the line marked *1 ISP* in Figure 1.

Path inflation occurs even in the simplest case, where both the source and destination are inside the same ISP, because ISP networks do not provide a direct *physical* link between all pairs of cities. The reasons behind this are largely economic – it is too costly to connect all cities directly. In Section 4.1 we expand on earlier work to map routers to their geographic locations [22, 27] and use this data to analyze path inflation due to the presence and absence of physical links. For instance, we observed single ISP paths that went from Belgium to Singapore via the USA, indicating the lack of more direct links connecting Europe and Asia.

Even when an intra-domain topology offers short paths between two cities, they may not be taken due to traffic engineering. The goal of intra-domain traffic engineering is to spread the load evenly among the links in the topology [10], but it may lead to longer paths. We show how to infer intra-domain routing policies in Section 4.2, and we analyze the impact of intra-domain policy on path inflation in Section 4.3.

2. ISP Peering: ISPs exchange traffic at select peering points in the network. Path inflation may result because packets are forced to transit one of these peering locations, which may or may not be "on the way" from the source to the destination. We analyze the inherent cost of crossing this ISP boundary in Section 5.1.

Peering policy refers to how an upstream ISP transfers traffic to a downstream ISP. The upstream ISP may not choose a peering point that optimizes latency. For instance, a common policy called *early-exit* chooses the peering point closest to the source of the packet (to minimize the cost incurred by the upstream ISP); the early exit may take the packets in a direction opposite to the ultimate destination. Since little is (publicly) known about peering policies, in Section 5.2 we develop techniques to measure the prevalence of *early-exit* and other peering policies, and in Section 5.3 we apply those techniques to measure the impact of policy on path inflation.

3. Inter-domain: Not all ISPs are directly connected to each other. As a result, a path from source to destination may traverse multiple intermediate ISPs. We analyze whether having to transit multiple ISPs leads to additional inflation in Section 6.1.

It is well-known that inter-domain routing may not select the best available paths [13, 26, 32]. Instead, the choice of paths is heavily influenced by the policies of the source, intermediate, and destination ISPs. We discuss common inter-domain policies, and study their impact in Section 6.2.

3. METHODOLOGY

This section describes our methodology for measuring topology and inferring routing policy. In summary, we collected a dataset of 19 million traces from 42 measurement sources over three days to discover 52,000 router IP addresses in the 65 ISPs we chose to study. ISP topologies are composed of backbones and POPs (point of presence) [27]. Each POP is a physical location (city) where the ISP houses a collection of routers, and the backbone connects these POPs. We assigned each router IP address to its respective POP and found 2,084 POPs in our dataset. In later sections, we analyze these POP-level topologies and traces. The raw data and maps are available at http://www.cs.washington.edu/research/networking/rocketfuel/.

3.1 Data Collection

As input for our analysis, we used traceroute data collected from 42 diverse PlanetLab vantage points [25], including sites in Australia, Canada, Denmark, Italy, New Zealand, Sweden, and the UK. This diversity is needed to get a complete picture of the network [17]. From these vantage points, we traced to all of the 125,000 prefixes in the BGP routing tables of RouteViews [20], which peers with sixty large ISPs. We used Scriptroute [28], a flexible network measurement facility, to collect the traces. Scriptroute enabled us to speed up the trace collection process while reducing the network load; it took us only 6 hours to collect traces from a vantage point to all 125,000 prefixes.

We collected trace data for three consecutive days, December 18, 19 and 20, 2002. Having traces from three days was useful for two reasons. First, vantage points that fail during trace collection on one of the days contribute traces on another day to complete the picture. Second, multiple datasets collected at different times enable us to filter out transient paths that may be observed due to instabilities such as failures. Since our focus in this work is to study inflation under stable conditions, we would like to filter out the impact of instabilities. Previous research has shown that transient routing events generally do not last longer than a day [8, 19, 24].

3.2 Choosing ISPs to study

To make this detailed study of routing policy feasible despite the size and complexity of the Internet, we carefully select individual ISPs for study. We used three criteria for picking ISPs: i) the ISP should be large enough to have interesting intra-domain and inter-domain choices to make; ii) the ISP should carry enough diverse traffic (be a transit provider) so that its topology and routing policy are easily observable using traceroutes (edge ISPs are much harder to observe); and iii) the set of ISPs should be diverse in size and geographic presence to show any resulting differences in topologies and policies.

We used the following informal methodology for selecting ISPs for detailed measurement. We first classified each ISP in the BGP tables to its *tier* [29]. Tiers represent how close to the Internet "core" an ISP is; tier-1 ISPs are the closest. We selected all 22 tier-1 ISPs and 32 high degree tier-2 ISPs. An ISP's degree is the number of neighboring ISPs in the BGP tables. Since the degree of an ISP is correlated to its size [34], high-degree ISPs are more likely to have interesting routing policies. To this mix, we added 11 tier-3 ISPs. The ISPs we study are listed in Table 1. The table shows that our selection process satisfied the three criteria listed above: the chosen ISPs have diverse degrees and are geographically diverse. After mapping these ISPs, we discovered that some have a very small or no backbone, providing another dimension in ISP diversity.

Even though we study a very small fraction of ISPs in the Internet, we believe that our chosen subset is useful for studying the impact of various factors on path inflation, because it includes most large providers in the network. Put together, these ISPs account for 40% of the singly homed, globally routed IP address space, implying that a significant fraction of Internet traffic traverses this set of ISPs. In fact, 97.7% of traces in our dataset traversed at least one of these 65 ISPs.

3.3 Extracting Topology

We now describe how we process traceroute data to recover ISP topologies. The first task is to identify which routers belong to each ISP using BGP and DNS. We use the BGP tables to distinguish the IP address space of different ISPs and then verify that the DNS name of the router matches the ISP's naming convention. Address space alone may blur the boundaries between ISPs and falsely place an ISP's customers and exchange points (a place where several ISPs peer) into the ISP's topology. For example, if Genuity runs an exchange point, the address of an interface on several non-Genuity routers may be part of Genuity's address space. DNS names provide confirmation that the IP address indeed belongs to the ISP.

We also map routers to geographical location (POP) by inference from their DNS names [22, 27]. For example, `gbr3-p10.st6wa.ip.att.net` is the name of an AT&T router interface in Seattle (st), Washington (wa). We build on Rocketfuel's DNS name decoding engine [27]. It takes a DNS name and the set of router name

ASN	Name	Tier	Dominant Presence	Degree
4637	Hong Kong Telecom	1	Asia-Pacific	199
6453	Teleglobe	1	Canada	162
8220	Colt	1	Europe	161
3320	DTAG	1	Europe	111
3300	Eqip	1	Europe	67
7176	Genuity-europe	1	Europe	90
5511	Open Transit	1	Europe	172
1299	Telia	1	Europe	256
7018	ATT	1	US	1490
3561	Cable Wireless	1	US	806
1	Genuity	1	US	622
3549	Global Crossing	1	US	585
4513	Globix	1	US	455
3356	Level3	1	US	539
4006	Netrail	1	US	14
209	Qwest	1	US	887
1239	Sprint	1	US	1735
701	UUNet	1	US	2569
2914	Verio	1	US	538
7911	Williams Comm	1	US	234
2828	XO	1	US	184
2497	IIJ	2	Asia-Pacific	165
4725	One Data Network	2	Asia-Pacific	63
4755	VSNL	2	Asia-Pacific	49
9942	Comindico	2	Australia	114
15290	ATT-Canada	2	Canada	81
577	Bellnexxia	2	Canada	88
6539	Group Telecom	2	Canada	155
3602	Sprint-Canada	2	Canada	53
852	Telus	2	Canada	82
2686	ATT-EMEA	2	Europe	62
5400	Concert	2	Europe	117
4589	Easynet	2	Europe	86
13129	GATel	2	Europe	53
9070	ITDNet	2	Europe	20
174	PSI	2	Europe	46
3257	Tiscali	2	Europe	326
702	UUNet-europe	2	Europe	587
5669	Vianw	2	Europe	36
15412	Flag Tel	2	International	32
1668	AOL	2	US	156
7170	ATT-Disc	2	US	60
11537	Abilene	2	US	86
11608	Accretive	2	US	124
2548	Allegiance	2	US	216
1785	Appliedtheory	2	US	86
6395	Broadwing	2	US	231
16631	Cogent	2	US	187
4544	Conxion	2	US	43
5650	Eli	2	US	172
4565	Epoch	2	US	84
1784	Gnaps	2	US	63
6939	Hurricane Electric	2	US	42
10910	Internap	2	US	113
101	PNW-GPOP	2	US	28
7132	SWBell	2	US	112
4323	TWTelecom	2	US	277
2687	ATT-AP	3	Asia-Pacific	24
7543	Singapore Telecom	3	Asia-Pacific	8
1221	Telstra	3	Australia	66
6509	CANet	3	Canada	16
6467	Espire	3	US	30
3967	Exodus	3	US	43
6461	MFN	3	US	498
3701	Nero	3	US	12
4600	Oregon-GPOP	3	US	4
12050	TLCT	3	US	21
3582	University Oregon	3	US	5

Table 1: ISPs studied, sorted by tier and their dominant regional presence. Many ISPs have some POPs outside this dominant operating region.

templates used by each ISP as input and proceeds in three steps: first, discover which ISP convention is used (AT&T), second, extract the location code (st6wa), and finally, map the location code to a city name (Seattle, WA).

Using this DNS-based tool, we reduce each router-level traceroute to a city-level path. This preserves routing policy because most traffic engineering decisions are made at the city (more precisely, POP) level [3, 31]. This step simplifies later analysis in three ways. First, it avoids the process of identifying the IP addresses that belong to the same router (alias resolution) and hence eliminates a potential source of error. Second, measuring a nearly complete backbone topology [2] is much easier than obtaining a complete router-level topology, requiring fewer sources [2, 17, 27]. Inferring policy from an incomplete topology is much more difficult. Third, the city-level topology has fewer nodes (2,084 versus 52,000), simplifying routing policy inference and analysis.

We took great care to ensure that our location mapping was complete and correct. While processing traces, our analysis flags names that match the router name pattern of ISP backbone routers but have undefined locations. We manually populate the tables of location codes based on this output until complete. Not all routers have location codes (Sprint customer routers, for instance); we infer the location of such routers from that of their neighbors [27]. A problem with trusting DNS names for router location is that routers may be improperly assigned to locations. The DNS name may be incorrect or stale, or the location inferred may be incorrect. For example, one IP address had the location tag "ALT," which is the three-letter airport abbreviation for Alenquer, Brazil. In this case, the router was in Atlanta, and was simply "ATL" mis-typed. We discover bad location matches by flagging physically impossible link latencies, limited by speed of light in fiber (in this example, Georgia to Brazil in 1 ms). We then disregard the trace until the router location inference has been corrected.

Large-scale, traceroute-based studies such as this one require an approach to filtering false links and paths that are observed during routing changes. False links may arise from routing changes during trace collection, since TTL-based path discovery is not atomic. We remove a link if it does not obey the speed-of-light criterion or if it is rarely used to carry traffic in our traces. For example, we would remove a direct link from Sydney to New York if, most of the time, the path between Sydney and New York traverses Los Angeles. It is important to remove these spurious links because they often represent a false, low latency, "non-stop" path through the network.

3.4 Inferring Policy

We study routing policy at three levels: intra-domain, peering, and inter-domain path selection. Our techniques for extracting these policies are different, and we defer the discussion of these techniques to later sections. However, each layer of our analysis builds on models validated at lower layers, so the overall accuracy of our results depends on the accuracy of previous steps.

3.5 Studying Path Inflation

We analyze the impact of topology and routing policy on path inflation between pairs of ISP cities. While aggregating statistics, we assume that all city pairs are equal regardless of their size or the amount of traffic they carry. However, some normalization happens automatically. In an ISP network, important "hub" cities are often relied upon by many smaller "spoke" cities to get to the outside world; hence, any path inflation between a pair of important cities impacts many city pairs. As a result, path inflation between two important cities impacts the aggregate measures more than inflation between smaller cities.

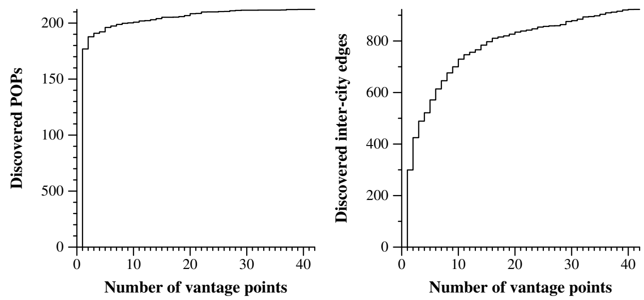

Figure 2: The incremental benefit of additional vantage points for mapping the network. The first vantage point finds the vast majority of nodes in the network. However, it takes several vantage points to capture most of the edges, and additional vantage points continue to contribute information.

While we focus on the inflation between city pairs, the absence of data describing the amount of traffic carried prevents us from extending our analysis from the fraction of paths to the fraction of packets that experience significant inflation. In particular, if we should find that paths between some fraction of POP-pairs suffer heavy inflation, that suggests but does not imply that a similar fraction of traffic will suffer inflation.

We compute link latencies using the geographic distance between the inferred location of adjacent routers. Except for networks that use circuit-switching, geographic distance correlates well with the minimum network delay [22]. This methodology is more robust than extracting link latencies from traceroute data, as they can be contaminated by queueing and unknown asymmetric reverse path delay. However, this methodology underestimates the latencies of links that do not take a direct path.

3.6 Completeness of the Dataset

Because our study is trace-driven, our results are limited by our choice of vantage points and ISPs. We might miss POPs and edges that belong to the ISPs we study. We conducted two simple analyses to assess the completeness of our dataset with respect to these ISPs. First, we compared the number of ISP pairs that peered in our traceroutes to the number that peered in the BGP tables. We found this fraction to be 71%. Second, we analyzed the amount of new information that each vantage point added. Figure 2 shows the number of POPs and city-level edges discovered as a function of the number of vantage points. Although nearly all of the POPs in our ISPs are captured with just a few vantage points, it is more difficult to capture all of the edges. We are well past the knee of the curve, however, and each additional vantage point contributes little new information. From these results we believe our topologies are sufficiently complete for our task of assessing path inflation.

4. INTRA-DOMAIN FACTORS

In this section, we first study the impact of intra-domain topology on path inflation, then describe our technique for inferring intra-domain policy, and finally analyze its impact on path inflation.

4.1 Impact of Topology

We study the impact of ISP topology by measuring path inflation along the shortest-latency path through the network. We compare this shortest-latency path to a hypothetical direct link between the two cities. The direct link "as the crow flies" serves as an ISP-independent reference point, and this comparison serves both to estimate how well-connected ISP topologies are and to put measurements of the inflation due to policy in context.

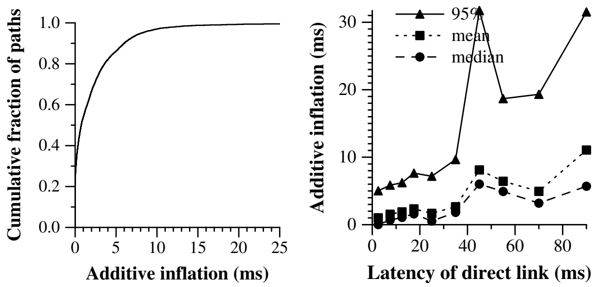

Figure 3: Path inflation due to intra-domain topological constraints. The left graph shows the CDF of additive inflation, and the right one shows the median and 95th percentile inflation as a function of latency of a hypothetical direct link.

We exclude ISPs that appear to use virtual circuit technologies such as MPLS. This prevents us from underestimating path inflation, as these topologies have many IP-level edges without a corresponding physical link. We identified six such networks by their unusually high edge to node ratio. We also exclude city pairs in the ISP that were never observed in the same trace. This prevents us from overestimating the inflation caused by topology if our traces missed a link that would provide a shorter path between the cities.

The graph on the left of Figure 3 shows the cumulative distribution function (CDF) of additive inflation. Most paths are not inflated by much, pointing to the well-connectedness of most ISPs. The median, mean, and 95th percentile inflation were 2, 3, and 10 ms, respectively. As explained below, the high inflation points correspond to city pairs on different continents.

The graph on the right characterizes inflation as a function of the reference distance between city pairs. We map the reference distance into bins – distances less than 20 ms are mapped to 5 ms bins (0–5, 5–10, 10–15, and 15–20); distances between 20 and 60 ms are mapped to 10 ms bins (20–30, 30–40, 40–50, 50–60); and longer distances are mapped to 20 ms bins. We use this binning in all similar graphs in this paper. It helps us to study the variation of inflation with the reference cost, as well as provide a sense of relative inflation. We found that the results are robust to bin sizes as long as the bins are small enough to expose effects that are present in a small latency range, and large enough to have sufficient data points to permit meaningful statistical measures.

The graph plots the median, mean, and 95th percentile of data points in each bin. We see that paths between most pairs of geographically close cities are inflated only by a small amount, indicating that these cities are usually connected by short paths. The jump around 40 ms represents cities that are on different continents; such cities are forced to reach each other by one of a few intercontinental links that exist in the intra-domain topology. The right end of the graph represents city pairs in Europe and Asia/Australia that have to traverse the USA. Since we filtered out city pairs that were not seen together in a trace, this is not an artifact of most of our vantage points being in the USA; our data contained intra-domain paths that went from Europe to Asia through the USA, observed from European vantage points.

Topological inflation among the tier-1 ISPs was higher compared to the other ISPs; the mean inflation was 4 ms for the former and 2 ms for the latter. This is largely a result of the tier-1 ISPs having more POPs over a larger geographic area, which makes it more difficult for them to connect all pairs with a good set of links. We observed no significant correlation between topological inflation and the ISP's geographic presence, suggesting that ISPs use similar topological designs on various continents.

4.2 Inferring Intra-domain Routing Policy

We next turn to the task of inferring intra-domain routing policy. We discovered that many paths observed in our dataset defied simple models based on hop count and link latency. Moreover, we did not have access to other factors, such as link capacity, that may influence intra-domain routing policy. Hence, instead of trying to guess the policy intent of each individual ISP, we sought a universally applicable model of intra-domain routing. We assumed that the routing is weighted shortest path with some unknown set of edge weights. Since we are dealing with POP-level topologies, these weights are different from the router-level link weights that ISPs use, but they serve a similar purpose. In this section, we outline a constraint-based approach to infer these weights; the approach is described in detail in our extended abstract [18].

Our approach takes the intra-domain topology and the set of paths observed over it as input, yielding edge weights that are consistent with the observed paths. The approach is based on the observation that the weight of an observed path (the sum of the edge weights) must be less than or equal to that of any alternate path between the same pair of nodes; otherwise, that path would have been preferred. This observation is translated into a set of constraints. For instance, if a path ABC was observed between A and C, and $ADEC$ is an alternate path, we set up the constraint $w_{AB} + w_{BC} \leq w_{AD} + w_{DE} + w_{EC}$. Similar constraints are set up for all observed and alternate paths, and the solution to the constraint system yields a set of edge weights that model observed routing.

Some paths in the dataset may have been observed due to transient events such as failures. In the basic scheme described above, such paths may lead to an inconsistent constraint system. We use constraint hierarchies [4] to account for this possibility. We associate an error variable with every observed path and set up constraints such that the error variable represents the weight of the path above the least weight path. In the example above, the constraint becomes $w_{AB} + w_{BC} - e_{ABC} \leq w_{AD} + w_{DE} + w_{EC}$.

Since the set of weights that model a given routing pattern is not unique, we also include constraints that minimize the deviation of the weight of an edge from its latency. The corresponding constraint is $w_{AB} - |e'_{AB}| = lat_{AB}$. It assumes that the latency and weight of an edge are related, and all things being equal, the weight increases with latency. This extension was not present in the previous paper [18].

We solve the constraint system while minimizing the sum of prioritized error variables [21]. The priority of a path's error variable is the number of times the path was seen, which means that transient paths are given very low priority. Edge error variables have very low priority to favor agreement with observed routing over similarity to link latency. Of the multiple weight settings that describe routing, we obtain the one that reflects the best correlation of weights and latencies.

We evaluated our inferred weights for their ability to characterize intra-domain routing in comparison to a pure latency-based model, in which an edge's latency is used as its weight. The left graph in Figure 4 shows, for each ISP, the percentage of observed paths that were least weight compared to the percentage that were shortest latency. The ISPs are sorted based on the success of weights in describing their routing. For many ISPs, weights describe routing much more closely.

The inferred weights not only fit the observed paths well, they also predict paths between cities for which no path was observed. We evaluated this by using half of our dataset to infer weights and analyzing how well these weights describe the complete dataset. To halve the dataset, we randomly removed all the paths between

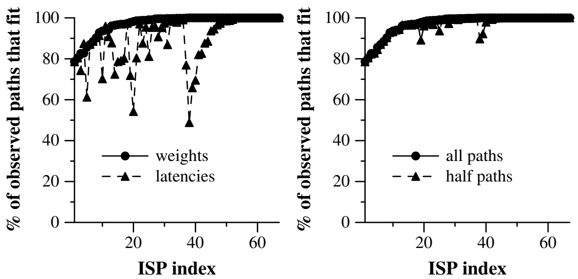

Figure 4: The success of the inferred weights in describing the observed routing. Each point along the x-axis corresponds to an ISP. The y-axis values depict the percentage of observed paths that were least cost. The left graph compares weights with latencies, and the right graph compares the weights inferred using the complete dataset with those inferred using half of it.

half of the city pairs. Removing individual paths would have just reduced the count of observations. The right graph in Figure 4 compares the success of the weights in describing the complete dataset when inferred using the complete dataset versus using only half of it. The weights inferred using half of the dataset predict routing nearly as well. We conclude that the inferred weights provide us with a concise and accurate model of ISP routing that is predictive beyond the paths present in the dataset used to infer it.

4.3 Impact of Intra-domain Routing Policy

We now analyze how intra-domain policy contributes to path inflation by measuring the inflation of least weight paths over shortest latency (geographic distance) paths. As in Section 4.1, we exclude networks that employ virtual circuit technologies and city pairs not observed together in a trace.

Figure 5 shows the results of the analysis. The overall median, mean, and 95th percentile were 0, 1, and 4 ms. The CDF on the left shows that 80% of the intra-domain paths are not inflated at all. This suggests that intra-domain traffic engineering is not inconsistent with latency-sensitive routing; ISPs try to minimize latency while tweaking weights to balance traffic across backbone links. The graph on the right plots inflation as a function of the shortest path latency using the binning described in Section 4.1. (The cost of the shortest path is used for binning.) Additive inflation rises as the shortest path latency increases. This makes intuitive sense, as there are more acceptable paths between these cities to pick from, allowing more opportunity for traffic engineering. Closer investigation of paths beyond 50 ms revealed that most appear to be caused by attempts to balance load across two or more transcontinental links. The decrease towards the right end of the graph is an artifact of having very few city pairs in that bin, most of which take an almost straight path through the topology.

As with topology, we found differences between tier-1 and other networks, but not between ISPs in different parts of the world. The inflation from intra-domain routing policy was higher for tier-1 ISPs; the mean was 1.2 ms compared to 0.3 ms for the other networks. This is most likely because tier-1 networks are bigger and have more redundant paths that allow traffic engineering.

5. PEERING FACTORS

In this section, we study the additional inflation that occurs when the source and destination connect to adjacent ISPs. First, we consider the peering topology – the union of all peering points between two ISPs. Geographically spread peering links provide more direct paths. Second, we consider the peering policy – the selection of which peering link to use to reach a destination.

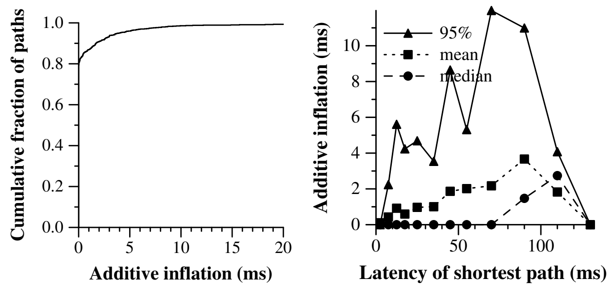

Figure 5: Path inflation due to intra-domain routing policy, when compared to shortest latency paths. The left graph shows the CDF of additive inflation, and the right one shows additive inflation as function of the latency of the shortest path.

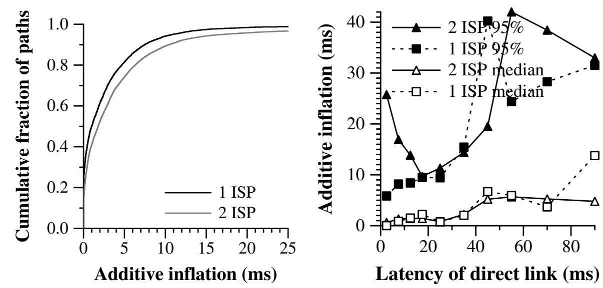

Figure 6: The cost of crossing an ISP boundary compared to intra-domain paths. The left graph shows the CDF of additive inflation over a hypothetical direct link between the city pair for paths that cross an ISP boundary (2 ISP) and intra-domain paths (1 ISP). The right graph shows the variation of median and 95th percentile inflation with the baseline hypothetical distance between the city pair.

5.1 Impact of Peering Topology

We measure the impact of peering topology by comparing paths that traverse two ISPs to those that stay within the same ISP. To isolate peering topology, we select optimal paths that cross two ISPs – the intra-domain portions use least-weight paths from the last section, but the peering link is chosen to minimize overall path latency. Our traceroutes, however, may not have exposed some optimal peering links that exist in the topology. To limit the resulting over-estimation of path inflation due to topology, we exclude cities in the upstream ISP from which we observed no path to the downstream ISP. Although the true optimal peering link may not be known, this exclusion ensures that at least chosen peering links are present.

The left graph of Figure 6 shows the CDFs of inflation compared to a hypothetical direct link for both intra-domain (1 ISP) paths and paths that cross a single ISP boundary (2 ISP). The difference between the two is noticeable but small, suggesting the presence of many peering links between adjacent ISPs. More interesting effects appear in the right plot, which shows the variation of inflation with the distance between source and destination. Nearby city pairs can have significant inflation if packets must travel from the source to a peering point and back to the destination.

Inflation from peering topology is least between networks with more peering points, such as tier-1 ISPs and networks in the same continent. Crossing boundaries between two tier-1 ISPs causes less inflation than when a smaller network is involved; the mean inflation compared to a hypothetical direct link was 4 ms between tier-1 ISPs and 7 ms otherwise. Crossing boundaries between two ISPs in the same continent is less costly than that between ISPs located

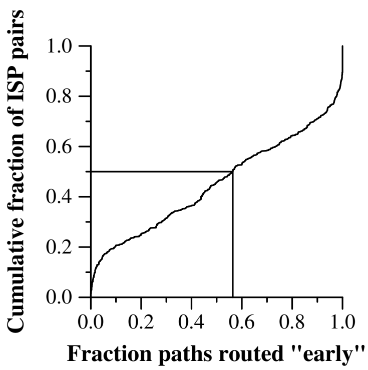

Figure 7: The prevalence of the early-exit routing policy. For those ISP pairs where we observed more than one peering link, this graph shows the fraction of paths we observed that were routed to the earliest exit. The median is 57%, meaning that most ISP pairs choose the early exit most of the time.

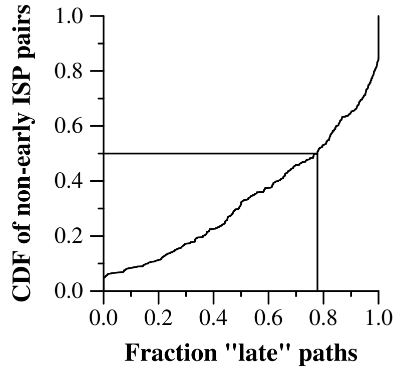

Figure 8: Of those ISP pairs that choose paths that are not "early," this graph shows the fraction of paths that are "helped" along toward the destination. That is, the first ISP carries the traffic closer to the destination than necessary, which implies some degree of cooperation and might be termed "late-exit." When the fraction of "helped" paths is small, another policy, such as "load-balancing" may be a more likely explanation.

in different continents (a mean of 5 ms in the former, and 9 ms in the latter).

In summary, we found that the peering topology does not inflate paths significantly. We next study the policies that govern the selection of peering points, then quantify how successful those policies are in finding short paths.

5.2 Characterizing Peering Policy

Since the peering topology between ISPs appears to be sufficient to support paths with little inflation in latency, the next potential source of inflation is routing policy over these links. For example, path inflation occurs when the selected peering point is to the east, while the destination and the optimal peering point are to the west.

The first step in understanding the inflation of paths caused by peering policy is understanding which policies are used and why. Common policies include *early-exit* and *late-exit*.[1] In *early-exit* routing, the upstream ISP uses a fixed peering point per source, usually the one closest to the source, independent of the destination. *Late-exit* is marked by the upstream ISP carrying the packets further than it needs to reach the closest peering point and towards the destination of the packet. In the extreme, this policy is source-independent, as the chosen peering point depends only on the destination. *Late-exit* requires cooperation among ISPs, because the upstream ISP cannot by itself know where the packet is destined inside the downstream ISP. That is, the downstream ISP must publish information (most likely via MED attributes in BGP) about the cost to reach destinations from each peering point, while the upstream cooperates by carrying traffic accordingly. We also uncover a third peering policy, that of *load-balancing*, in which an upstream ISP balances load across multiple peering links.

We classify routing patterns using the simplest explanation consistent with our traces. We discard ISP pairs for which we observe only a single peering point. The rest are classified as *early* if only one peering point is seen from each ingress. What remain are peerings that show "engineering" (assuming *early-exit* to be the default), so the first question to ask is whether this engineering is consistent with *late-exit*.

To discover whether the engineering appears cooperative in preventing latency inflation, we identify as *late* those paths that are delivered by the upstream ISP to a peering point closer to the destination in the downstream ISP. That is, the path length in the downstream ISP from peering point to destination is less than from the early exit to the destination. We use this metric to classify policies into three broad classes: $i)$ *Late exit, often* represents a pattern of late exit for most paths; $ii)$ *Late exit, sometimes* represents a selective pattern of late exit for a minority of paths; $iii)$ *Engineered, but not late* represents a pattern where the downstream ISP carries traffic over longer paths, perhaps as part of a "load-balancing" strategy.

In Figure 7, we show that while half of all ISP pairs route at least half of their prefixes using early-exit, many do not, hinting at widespread cooperation among ISPs. Additionally, there is significant variation in the degree of path engineering. To the right of the graph, 20–30% of ISP pairs appear to use only early-exit; to the left, 10–20% route with primarily *late-exit;* and in between is a wide range.

For those ISPs that do not always use early-exit, we next consider how often the first ISP carries traffic closer to the destination. In Figure 8, we show a cumulative distribution of the fraction of *late-exit* paths. The cooperation among ISPs is apparent in this graph, too: most ISP pairs that do not use *early-exit* route over 75% of paths closer to the destination. For instance, Qwest routes almost all prefixes destined to Genuity using a *late-exit* that is independent of the source. A few ISP pairs route only a few paths close to the destination: suggesting either micro-engineering of a few prefixes or a limited load balancing policy.

In Figure 9, we summarize the apparent peering policy of tier-1 ISPs we inferred from our traces. We classify each pair of ISPs once for each direction, since peering policy need not be (and often is not) symmetric. Those pairs labelled *Late-exit, often* deliver more paths late (closer to the destination) than early. Those labelled *Late-exit, sometimes* deliver at least 5% of paths late. *Source-dependent early-exit* refers to those ISP pairs for which fewer than 5% of paths are routed specially. A threshold is needed to be robust to anomalous routes during failures. *Single peering seen* have a single peering link in one location, often connecting ISPs on different continents. Policy has no effect for these ISP pairs. *Engineered, but not late* refers to those ISPs with many paths that are not early, but frequently (at least 1/3 of the time) are not carried closer to the destination. Finally, some ISP pairs appear to peer based on RouteViews [20] BGP table archives but were not observed in our dataset.

[1] There are no standard definitions for peering policies. Our characterization is based on the behaviors we observed in our dataset, rather than informally used terms.

From / To	ATT	ATT-Disc	Cable Wireless	Colt	DTAG	Eqip	Genuity	Genuity-europe	Global Crossing	Globix	Hong Kong Telecom	Level3	Netrail	Open Transit	Qwest	Sprint	Teleglobe	Telia	UUNet	Verio	Williams Comm	XO
ATT		!!	!		!	x	●		●	●		!	○	!	!!	x	●	!	x	●	!	x
ATT-Disc	○						?		?	?					x					?	?	?
Cable Wireless	!!			x	!	○	!		x			●	●	x	!!	●	!!	!	●	●	●	!
Colt			?				?			?	?	?		?						○		
DTAG	?		○	●			○	?	?	?	?	!		○	?	?	?	○		○	?	?
Eqip	?		?				?		?	?	?			?	○	?			?	?	?	?
Genuity	x	?	●	○	●	●		!	x	x	!!	●	○	x	!!	●	●	x	●	●	●	●
Genuity-europe					?		●		?	?	○	?		?		○						
Global Crossing	!!	!!	!!	○	x	○	!!	!		○	○	●	○	!	●	!!	!!	○	!	!!	○	x
Globix	?	?		?	?	?	?	?	?		?	?		?		○			x	?	?	?
Hong Kong Telecom			○	○	?	●	●	○	●	?		?		?	○	○				○	?	?
Level3	x		!	●	!!	○	!!	○	●	●	●		○	x	!	●	●	!	!!	!	x	●
Netrail	○		●		○		○		○		○	○		○	●	○			?			?
Open Transit	?		○	○	?	○	○	?	?	?	?	!	?		○	○	?	●	○	?		?
Qwest	x	!	x	○	○	○	!!		!		!	x	●	x		x	x	x	!!	x	x	x
Sprint	x		x	○	●	○	x	?	●	!	!	x		x	●		x	!!	!!	x	!	!!
Teleglobe	!!		●		○		●		●			!!		!!	x	!!		○	!!	x		?
Telia	!!		!!	○	!!	x	!!	●	!!		○	!!	○	!!	!!	!!	!		!!	!!	!!	
UUNet	●		x			!	●		●	!	○	●	○	●	!	x	x	●		●	!	●
Verio	●	○	x	!!	!!	x	●	○	!	x	!	!		x	●	●	x	●	!		x	x
Williams Comm	?	?	!!		?	○	?		○	?	○	○			●	●		○	!!	○		?
XO	●	?	!		?	?	●		○	?	○	○	?	○	○	●	?		!	○	○	

!!	"Late exit," often	(15%)	●	Source-dependent "early-exit"	(19%)	x	Engineered, but not "late"	(13%)
!	"Late-exit," sometimes	(10%)	○	Single peering seen	(42%)	?	Peering in RouteViews but unseen	

Figure 9: Summary of observed peering relationships between tier-1 ISPs. Both early- and late-exit policies are common, and not simply the work of a few ISPs. Peering policies vary widely, even between neighbors of the same ISP.

Some ISPs, such as Telia and Global Crossing, appear to use late exit more often than others. In all, this figure demonstrates the diversity of pairwise peering policies, and that "early-exit" rarely sums up routing policy between tier-1 ISPs.

We found asymmetry in the peering policies used in general and associated with the tier of the ISP. Of those ISP pairs we were able to characterize as early or late in both directions, we found that the likelihood of a late exit policy is 17%, but the likelihood that an ISP reciprocates late exit is only 8%. We also found asymmetry in the relationship between tier-1 and tier-2 ISPs. 50% (82/178) of the tier-1 to tier-2 peerings showed a late-exit policy compared to only 36% (32/88) of the tier-2 to tier-1 peerings. We found no major differences in the policies used in different parts of the world.

5.2.1 A Case Study: AT&T and Sprint in San Francisco

While most peerings can be classified cleanly as early or late, there are some paths that traverse a peering link that is neither closest to the source nor closer to the destination. We looked in detail at a particularly pronounced case, between AT&T and Sprint in San Francisco, California. We discovered that only 38% of 14,649 paths took the early-exit peering in San Francisco. The other 61% of paths were sent to Seattle, Washington, 1092 kilometers north. Many paths diverted to Seattle returned to San Francisco, making "late-exit" an unlikely explanation.

We first evaluate how this policy inflates path latencies and then examine a potential explanation. Figure 10 shows the CDF of additive inflation in paths leaving AT&T's San Francisco POP for destinations in Sprint. The inflation is relative to optimal exit paths, computed as in Section 5.1. We compare the inflation in measured paths to the inflation that would result from an early-exit policy. While the early-exit policy would cause little path inflation, the observed routing sees at least 5 ms inflation on most paths.

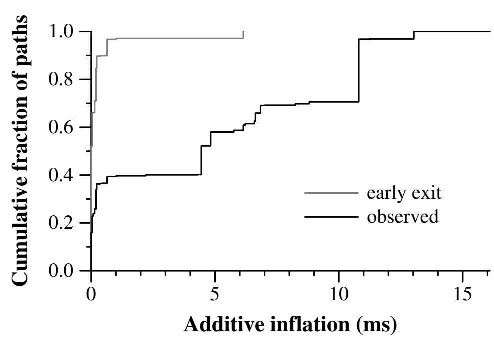

Figure 10: The impact of AT&T's unilateral load balancing on path inflation. The graph plots the CDF of additive inflation in early-exit and observed paths when compared to optimal-exit paths from the San Francisco peering point.

We wondered why AT&T would deviate from simple early-exit and make paths longer.[2] We postulated that AT&T engineers were diverting traffic away from a heavily loaded peering link and set out to corroborate this assumption by measuring congestion on the link. Between January, 16 and 24, 2003, we sent a series of TTL-limited probes from PlanetLab sites in Seattle and Berkeley to mea-

[2]We dismissed sinister intentions, such as causing Sprint to carry the traffic farther in its network.

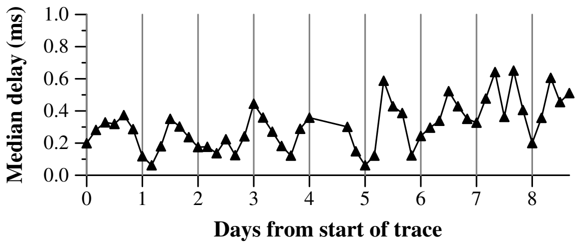

Figure 11: The median delay on the San Francisco peering between AT&T and Sprint with 4 hour averaging. The vertical bars represent midnight on each day.

sure the queueing delay at both Seattle and San Francisco peering links. The probes were sent in triplets, every 50 seconds to avoid synchronization with a router's periodic maintenance tasks (usually 60 seconds [23]). These probes were meant to expire at the two routers immediately before the link and at the router immediately after the link. To guard against routing changes, we verified that the responses to each probe came from the same router and the return TTL was consistent. For a given link, variation in RTT to the remote end without variation in RTT to the near end represents queuing delay variation and is hence an indication of load.

We observed no queueing delay variation on the links before both peering links and on the Seattle peering link. The San Francisco peering link, however, appeared loaded. Figure 11 shows the measured queueing delay at the link, with each dot representing the median delay over a four hour period.[3] The vertical bars correspond to midnight, Pacific time, on each day. That the measured queueing delay decreases almost every night strongly suggests that we are measuring a quantity dependent on the load at the link. (This variation implies that ICMP generation delay at the router is not a factor, consistent with [14].) The third day in the trace was a Saturday; interestingly, the load increased on Saturday and Sunday nights. Our measurement point failed on the fifth day, removing some Monday morning data points.

In this paper, we use path inflation to infer the presence and character of engineering in the network. In this case study, we found that the seemingly abysmal routing policy of sending traffic to Seattle was instead a result of trying to avoid a congested peering link. Of course, the ideal solution is to add capacity or at least to load balance traffic in a topology-sensitive manner to prevent inflation. Topology-sensitive load balancing is difficult in BGP, which we discuss in Section 7.

5.3 Impact of Peering Policies

We next consider the path inflation of alternate peering policies. We focus first on early- and late-exit policies, then compare to a latency-optimized policy.

5.3.1 Early- vs. Late-Exit

We found that there is little difference in latency between the early and late exit routing strategies. Figure 12 shows the CDF of inflation caused by early- and late-exit policies from Qwest to Genuity. Qwest implements a late-exit policy towards Genuity, but in terms of path latency, early would have been just as good. This is not surprising, because the paths taken by late-exit are usually just the reverse of paths taken by early-exit. Although these policies differ for practical purposes, for the rest of this paper, we analyze peering policies using early-exit as representative of both early and late exit.

[3]With an OC-3 link (155 Mbps), a 0.5 ms median queueing delay implies a median queue size of roughly 20 packets.

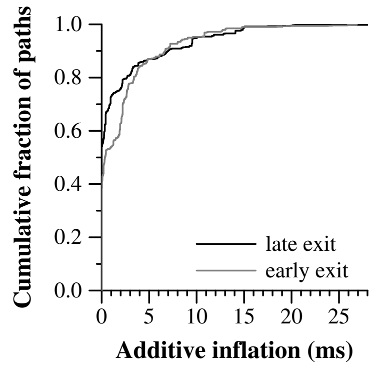

Figure 12: The inflation caused due to early- and late-exit policies when going from Qwest to Genuity.

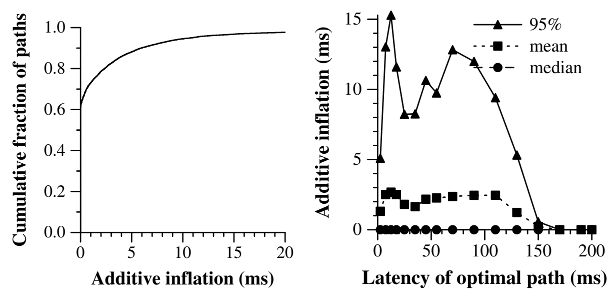

Figure 13: The path inflation due to *early-exit* compared to optimal exit.

5.3.2 Early- vs. Optimal-Exit

Figure 13 compares the inflation caused by using early-exit routing relative to an ideal, optimal exit policy. ISP pairs with only one peering point in the traces have been excluded. The left graph shows the CDF of additive inflation for early relative to ideal. Over 30% of the paths experience inflation, and the top 5% of the paths suffer an inflation of more than 12 ms. The graph on the right shows the inflation as a function of the latency of the optimal exit path. It shows that paths between cities close to each other can be highly inflated, compared to the small distance between them. Unlike previous graphs, the additive inflation does not increase with the latency of the optimal path. At the right end of the graph, many paths face a choice between two, closely located, trans-continental links. We did not observe any significant dependence of our results on the size of ISPs.

In summary, our results show that peering policies often cause path inflation, and these policies sometimes lead to highly inflated paths. This means that they consume more network resources, probably for both ISPs, than an optimal exit policy.

5.3.3 More Peering Points vs. Optimal-Exit

We next consider adding additional peering points between ISPs. The intent is to determine whether an "optimal exit" policy might have the same effect as adding peering points. We focus on the topology of Sprint and AT&T, though similar results were obtained using other US ISP pairs. We found 11 peering points between Sprint and AT&T spread all over the country. To compare optimal- and early-exit path inflation on topologies with fewer peering points, we repeat our inflation analysis over topologies with randomly chosen subsets of these 11 peering links, averaging over 10 topologies for each degree of peering.

Figure 14 shows the results of the experiment. It plots the median and the 95th percentile of additive inflation over the latency

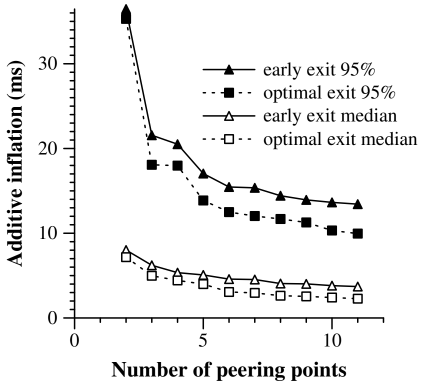

Figure 14: The variation of additive inflation with the number of peering points between Sprint and AT&T. Additive inflation is measured with respect to the latency of a hypothetical direct link between the source and the destination cities.

of a hypothetical direct link between the source and the destination cities. With just two or three peering points, the inflation is high, as all traffic is constrained to traverse these cities. The inflation decreases quickly as the number of peering points increases.

There is a 3 ms gap between the 95th percentiles of optimal- and early-exit routes. Somewhat surprisingly, the performance difference between early-exit and optimal-exit peering policies does not decrease with more peering points. That is, an optimal-exit peering policy continues to offer lower latency paths. This graph can also be read horizontally to show that the latency reduction offered by more peering points can instead be achieved using optimal-exit.

6. INTER-DOMAIN FACTORS

In this section, we describe the path inflation along paths that traverse multiple ISPs. ISPs choose which other ISPs to connect with; these organizational relationships form the AS-graph that defines the paths that exist in the network. Over this topology, ISPs make routing policy decisions that determine which paths are taken. This path selection is constrained by business relationships that may forbid good paths that exist in the topology. In this section, we want to determine whether it is the topology, the protocol, or the policies that create the inter-domain path inflation in Figure 1.

To estimate the path inflation contributed by inter-domain topology and policies, we construct an abstract graph where nodes represent POPs and edges represent both the early-exit paths between ISPs from Section 5 and the least-weight paths between POPs in the same ISP from Section 4. In this section, we compute shortest paths over this graph, first minimizing latency to show the effect of topology, then minimizing AS-hops to compare policies. We ensure that these shortest paths respect intra-domain and peering policies by using the intra-domain edge only at the end of the path. (Otherwise, optimal exit could be synthesized by composing intra-domain edges in the middle of the path.) There are 100,799 edges connecting 2,084 nodes in this graph. In synthesizing paths between POPs in this section, we do not filter out node pairs that were not observed in the traces, which may bias our results towards estimating more path inflation than actually exists. In addition, with more ISPs in a path, the mean distance between city pairs is increased, implying that caution is reasonable in comparing the quantitative results in this section to the results in earlier sections. However, our results are consistent with previous studies of measured path inflation [26, 32].

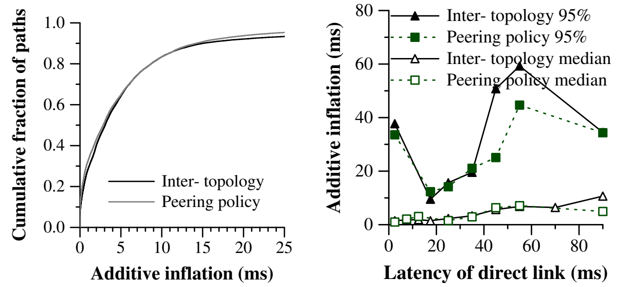

Figure 15: Path inflation due to inter-domain topology relative to a hypothetical direct link. Each ISP chooses a lowest-latency path to reach destinations. The peering policy line represents the inflation caused by the factors in Sections 4 and 5.

6.1 Impact of AS-graph

In this subsection, we measure the inflation of the lowest latency, intra-domain, and peering policy compliant path between POPs in different ISPs. In later subsections, we will constrain inter-domain path selection using common policies that do not minimize latency.

If the AS-level topology of the Internet was sparse, we would expect that paths might cross several ISPs to reach their destination, perhaps contributing significant path inflation. However, AS-paths through the Internet are generally short [16], and the subset of the network we study consists of tier-1 and well-connected tier-2 ISPs. As a result, we expect inter-domain topology to contribute only modest path inflation.

In Figure 15, we show the path inflation experienced along the lowest-latency AS path between any two POPs in the network. This represents the inflation when each ISP chooses the best (in terms of latency) next-hop AS without changing intra-domain routing or peering policy. On the same graph, we also plot the inflation between POPs in adjacent ISPs from Section 5.3. Although different samples of POP pairs are used for the two lines, and so should not be compared directly, crossing multiple ISPs does not appear to contribute significant additional path inflation.

6.2 Impact of Inter-domain Routing Policy

We now consider how well inter-domain routing policy chooses paths using this topology. There are two fundamental policies used by ISPs that represent common business arrangements [12, 13]:

No-valley Peers[4] and customers do not generally provide transit (carry traffic to other peers). Without paying customers, there is no incentive for an ISP to carry traffic.

Prefer-customer Paths through customers are preferred to those through peers, which are in turn preferred to those via providers. Paths through customers generally represent income; through providers, an expense; and through peers, no (or little) cost.

These rules have the potential to create path inflation, because they prohibit paths that exist in the topology from being used. We inferred the (provider-customer and peer-peer) relationships between ISPs using Gao's analysis [12]. We then simulate these policies in path selection over the topology; each ISP chooses and propagates legal no-valley paths, using the prefer-customer rule before breaking ties by shortest AS path, then lowest intra-domain cost. Among paths that conform to no-valley and prefer-customer policies, paths that traverse the fewest ISPs (shortest AS-path length) are usually

[4]The term *peer* is overloaded. Following common practice, in previous sections, we use it to denote an interconnection between ISPs, and here we use it to represent a commercial relationship between comparably-sized ISPs that exchange traffic without payment.

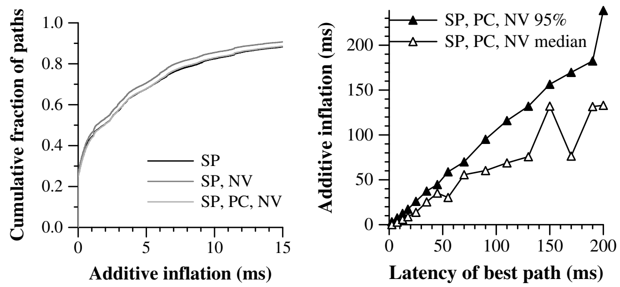

Figure 16: Path inflation due to common AS policies. With the latency of the lowest-latency path as a reference (the x-axis), we show the inflation when choosing shortest AS paths (SP), then choosing only valley-free paths (NV), then additionally preferring customer (PC) routes. At right, only the no-valley, prefer-customer inflation is shown; the other lines are similar.

chosen. As in current inter-domain routing, the latency of the path is not considered directly.

In Figure 16, we show the inflation caused by prefer-customer and no-valley policies, relative to the lowest-latency path characterized in Figure 15. Each of the series is quite similar because most of the inflation is caused simply by choosing the shortest AS path. The no-valley line on the CDF shows slightly less inflation – the rule prevents some bad paths – while adding prefer-customer chooses slightly more inflated paths. These differences do not appear to be significant; the graph primarily shows the inflation that results from shortest AS path routing. These results are consistent with the findings of Tangmunarunkit *et al.* [32], who evaluate these policies based on router hop counts. At right, we show that it is the long paths that are made longer. Unlike intra-domain paths shown in Figure 5, shortest AS paths do not minimize latency.

In summary, we find that inter-domain routing in the current Internet has a significant impact on path inflation with more than half of the paths being longer than the shortest path. However, policies such as no-valley and prefer-customer arising out of commercial concerns are not a contributing factor. Instead, most of the inflation comes from using shortest AS-path as the distance metric.

7. SUMMARY

Figure 17 shows the cumulative impact of all six factors on path inflation, computed with reference to a direct hypothetical link between the source and the destination. The inflation caused by a particular factor alone can be isolated by subtracting its inflation from the previous factor. The two biggest factors are inter-domain and peering policy. Interestingly, the median inflation caused by inter-domain policy (6.9–3.0 = 3.9 ms) is much more than the median inflation caused by intra-domain topology, which represents physical link constraints.

We argue that the observed inflation in paths due to peering and inter-domain policies is not the result of desired routing patterns but signifies a lack of good tools for ISPs to find better paths. Our results in Section 4.3 suggest that ISPs optimize their own paths for latency, as long as it does not cause congestion. In Section 5.2, we demonstrated widespread cooperation among ISPs to engineer better paths. Putting these results together, we argue that the given the right set of tools, ISPs would use them to reduce both congestion and path inflation.

At both the peering and inter-domain levels, there is an absence of mechanisms in BGP to enable better path selection. For the former, there is no convenient way to do optimal exit routing or topology-sensitive load-balancing across peering links. BGP MEDs are designed to assist in the selection of peering links; downstream ISPs propagate their preference for which link to use to the upstream ISP. The upstream ISP, however, has no semantically meaningful way to combine these preferences with their internal shortest-path lengths. One must either ignore MEDs (early-exit) or listen to them (late-exit).

		Median	Mean	95%
Intra-domain	Topology	1.0 ms	2.4 ms	8.4 ms
	Policy	1.4 ms	3.2 ms	11.5 ms
Peering	Topology	2.0 ms	5.0 ms	17.7 ms
	Policy	3.0 ms	6.5 ms	24.5 ms
Inter-domain	Topology	3.0 ms	7.3 ms	34.1 ms
	Policy	6.9 ms	13.9 ms	60.3 ms

Figure 17: Cumulative path inflation caused by each of the six factors, computed with reference to a hypothetical direct link.

Similarly, BGP does not propagate enough information to enable an informed AS path selection, and as a result, ISPs often use minimum AS-hop count paths. We showed in Section 6.2 that paths with minimum AS-hop count are often longer than needed. In practice, operators try to tune performance when the defaults are unacceptably bad. Even then, this involves cumbersome manual configuration, which is error-prone [19] and may be sub-optimal, as in the case of AT&T's San Francisco peering with Sprint.

We believe path selection can be significantly improved if effective mechanisms to achieve these goals were added to BGP. One possible mechanism would append geographic coordinates to route advertisements. These coordinates would denote where a route enters and exits each ISP in the path without publishing details of the topology of an ISP. Our measurements indicate that geography is a good indicator of latency for paths that lie within most ISPs. The upstream ISP could then choose AS paths and peering links to minimize latency and use location-sensitive load balancing for congested peering links.

8. RELATED WORK

Our work draws on and benefits from three separate lines of work – topology collection and analysis, policy inference, and studies of the impact of routing policies.

Various researchers have collected and analyzed Internet topology at different granularities, such as router-level [15, 7, 27] and AS-level [9, 6]. We use the techniques of router-level topology collection (traceroute) and router name-to-location mapping [27, 22] to collect POP-level topologies of 65 major ISPs and their interconnections. Extending previous work that studied topology properties, such as degree distribution [9, 35, 5] and resilience [35], we study how topology impacts path inflation.

Inference of routing policy to date has been limited to the inter-domain case [12, 29]. We add two key missing pieces to routing policy inference work: inference of intra-domain traffic engineering and inference of peering policy between a pair of ISPs.

Starting with the Detour project [26], much attention has been given to the impact of inter-domain routing policies on performance. Detour showed that there exist paths in the Internet that are better than those being used. While it measured the end-to-end impact of routing policies, we infer the topology and routing policies at various levels and use those results to isolate the effect of each level of routing on end-to-end performance.

Subramanian *et al.* [30] use geographic distances to study the "circuitousness" of network paths using a dataset of fifteen sources and six thousand destinations. Through analysis of the relative distance traveled in the upstream versus the downstream ISP, they hint at the prevalence of early-exit peering policy. We use a much larger data set and topology inference to expose the diversity and impact of peering policies.

The impact of inter-domain path selection has been studied in some detail. Tangmunarunkit *et al.* [33] studied inter-domain path selection and observed that paths are frequently chosen based on criteria other than the shortest AS path. Two studies then followed to attribute this AS-level path inflation to the no-valley and prefer-customer BGP policies. While Wang and Gao found significant inflation of AS-paths due to these policies [13], Tangmunarunkit *et al.* found router hop lengths did not increase significantly as a result [32]. Our study extends this work to show that path latency does not increase significantly due to no-valley and prefer-customer policies.

9. CONCLUSIONS

In this paper, we analyzed how topology and policy combine to contribute to path inflation at three levels in the Internet: intra-domain, peering, and inter-domain. While longer paths are not always harmful, they are often symptomatic of traffic engineering practices in the Internet.

Through a trace-driven analysis, we isolated the effects of the six causes of path inflation: topology and policy at all three levels. We collected city-level maps of 65 ISPs and their interconnections using traces from 42 vantage points to all globally-routed IP address prefixes. We also devised novel techniques to infer the intra-domain routing policy and the peering policy among pairs of ISPs. We used these models to evaluate observed and alternative routing policies in terms of path inflation.

We found that setting weights on intra-domain links does not significantly impact path inflation, suggesting that intra-domain traffic engineering is latency-sensitive. Our analysis exposed diverse peering policies, from late-exit to micro-engineered to load-balanced, and we provided the first empirical data on the prevalence of these policies. Although early-exit is used by most ISPs most of the time, most engineer at least some paths to accomplish policy goals. Unlike intra-domain policies, current peering policies do impact path inflation, with over 30% of the paths being longer than the shortest available path. We also found that inter-domain routing causes significant path inflation, with almost half the paths being inflated, not because of policies arising out of commercial concerns, but because shortest AS path is the default.

The latency sensitivity of intra-domain routing and the cooperation among ISPs to engineer better peering exits suggests that ISPs are willing to work together to achieve low latency routing. In view of this, we believe that Internet path inflation can be significantly reduced if effective mechanisms were made available to ISPs to engineer better peering point and AS-path selection.

10. ACKNOWLEDGEMENTS

We would like to thank David Wetherall for his help and advice throughout this project. Steve Gribble, Ken Yasuhara, Andy Collins, Richard Dunn, Krishna Gummadi, Intae Kang, Eric Lemar, Mingyan Li, Javier Salido, Radhakrishna Sampigethaya, Stefan Saroiu, Stefan Sigurdsson, Mike Swift, Gang Zhao, and the SIGCOMM reviewers also provided very helpful feedback on various drafts. Intel Research provided access to PlanetLab, and CAIDA provided Skitter data for a preliminary version of this study. This work was supported in part by Intel Corporation and by DARPA grant F30602-00-2-0565.

11. REFERENCES

[1] D. Anderson, H. Balakrishnan, M. F. Kaashoek, and R. Morris. Resilient overlay networks. In *SOSP*, 2002.

[2] P. Barford, A. Bestavros, J. Byers, and M. Crovella. On the marginal utility of network topology measurements. In *ACM SIGCOMM Internet Measurement Workshop*, 2001.

[3] S. Bhattacharyya, C. Diot, J. Jetcheva, and N. Taft. Pop-level and access-link-level traffic dynamics in a Tier-1 POP. In *ACM SIGCOMM Internet Measurement Workshop*, 2001.

[4] A. Borning, B. Freeman-Benson, and M. Wilson. Constraint hierarchies. *Lisp and Symbolic Computation*, 5(3), 1992.

[5] T. Bu and D. Towsley. On distinguishing between Internet power law topology generators. In *IEEE INFOCOM*, 2002.

[6] H. Chang, *et al.* On inferring AS-level connectivity from BGP routing tables. Tech. Rep. UM-CSE-TR-454-02, University of Michigan, 2002.

[7] k. claffy, T. E. Monk, and D. McRobb. Internet tomography. In *Nature*, 1999.

[8] M. Dahlin, B. Chandra, L. Gao, and A. Nayate. End-to-end WAN service availability. In *USITS*, 2001.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *ACM SIGCOMM*, 1999.

[10] A. Feldmann, *et al.* Netscope: Traffic engineering for IP networks. *IEEE Network Magazine*, 2000.

[11] P. Francis, *et al.* IDMaps: A global Internet host distance estimation service. *IEEE/ACM Transactions on Networking*, 2001.

[12] L. Gao. On inferring autonomous system relationships in the Internet. In *IEEE Global Internet Symposium*, 2000.

[13] L. Gao and F. Wang. The extent of AS path inflation by routing policies. In *IEEE Global Internet Symposium*, 2002.

[14] R. Govindan and V. Paxson. Estimating router ICMP generation delays. In *Passive & Active Measurement (PAM)*, 2002.

[15] R. Govindan and H. Tangmunarunkit. Heuristics for Internet map discovery. In *IEEE INFOCOM*, 2000.

[16] G. Huston. BGP statistics. http://bgp.potaroo.net/rv-index.html.

[17] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, 2003.

[18] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson. Inferring link weights using end-to-end measurements. In *ACM SIGCOMM Internet Measurement Workshop*, 2002.

[19] R. Mahajan, D. Wetherall, and T. Anderson. Understanding BGP misconfiguration. In *ACM SIGCOMM*, 2002.

[20] D. Meyer. Routeviews project. http://www.routeviews.org.

[21] K. G. Murty. *Linear Programing*. John Wiley & Sons, 1983.

[22] V. N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for Internet hosts. In *ACM SIGCOMM*, 2001.

[23] K. Papagiannaki, *et al.* Analysis of measured single-hop delay from an operational backbone network. In *IEEE INFOCOM*, 2002.

[24] V. Paxson. End-to-end routing behavior in the Internet. In *ACM SIGCOMM*, 1997.

[25] L. Peterson, T. Anderson, D. Culler, and T. Roscoe. A blueprint for introducing disruptive technology into the Internet. In *HotNets-I*, 2002.

[26] S. Savage, *et al.* The end-to-end effects of Internet path selection. In *ACM SIGCOMM*, 1999.

[27] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with Rocketfuel. In *ACM SIGCOMM*, 2002.

[28] N. Spring, D. Wetherall, and T. Anderson. Scriptroute: A public Internet measurement facility. In *USITS*, 2003.

[29] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet hierarchy from multiple vantage points. In *IEEE INFOCOM*, 2002.

[30] L. Subramanian, V. N. Padmanabhan, and R. H. Katz. Geographic properties of Internet routing. In *USENIX Annual Technical Conference*, 2002.

[31] N. Taft, S. Bhattacharyya, J. Jetcheva, and C. Diot. Understanding traffic dynamics at a backbone POP. In *SPIE ITCOM Workshop on Scalability and Traffic Control in IP Networks*, 2001.

[32] H. Tangmunarunkit, R. Govindan, and S. Shenker. Internet path inflation due to policy routing. In *SPIE ITCom*, 2001.

[33] H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin. The impact of routing policy on Internet paths. In *IEEE INFOCOM*, 2001.

[34] H. Tangmunarunkit, *et al.* Does AS size determine degree in AS topology? *ACM Computer Communication Review*, 2001.

[35] H. Tangmunarunkit, *et al.* Network topology generators: Degree-based vs structural. In *ACM SIGCOMM*, 2002.