The Impact of BGP Dynamics on Intra-Domain Traffic

Sharad Agarwal
CS Division,
University of California, Berkeley
sagarwal@cs.berkeley.edu

Chen-Nee Chuah
ECE Department,
University of California, Davis
chuah@ece.ucdavis.edu

Supratik Bhattacharyya
Sprint ATL,
Burlingame, CA, USA
supratik@sprintlabs.com

Christophe Diot
Intel Research,
Cambridge, UK
christophe.diot@intel.com

ABSTRACT

Recent work in network traffic matrix estimation has focused on generating router-to-router or PoP-to-PoP (Point-of-Presence) traffic matrices within an ISP backbone from network link load data. However, these estimation techniques have not considered the impact of inter-domain routing changes in BGP (Border Gateway Protocol). BGP routing changes have the potential to introduce significant errors in estimated traffic matrices by causing traffic shifts between egress routers or PoPs within a single backbone network. We present a methodology to correlate BGP routing table changes with packet traces in order to analyze how BGP dynamics affect traffic fan-out within a large "tier-1" network. Despite an average of 133 BGP routing updates per minute, we find that BGP routing changes do not cause more than 0.03% of ingress traffic to shift between egress PoPs. This limited impact is mostly due to the relative stability of network prefixes that receive the majority of traffic – 0.05% of BGP routing table changes affect intra-domain routes for prefixes that carry 80% of the traffic. Thus our work validates an important assumption underlying existing techniques for traffic matrix estimation in large IP networks.

Categories and Subject Descriptors

C.2.3 [**Computer Systems Organization**]: Computer-Communication NetworksNetwork Operations; C.4 [**Computer Systems Organization**]: Performance of Systems

General Terms

Measurement, Performance, Algorithms, Management, Reliability

Keywords

Traffic Matrix, Traffic Engineering, Traffic Analysis, BGP

1. INTRODUCTION

The Internet is an interconnection of separately administered networks called Autonomous Systems or ASes. Each AS is a closed network of end hosts, routers and links, typically running an intra-domain routing protocol or IGP (Interior Gateway Protocol) such as IS-IS (Intermediate System to Intermediate System) [1] or OSPF (Open Shortest Path First) [2]. The IGP determines how a network entity (end host or router) inside the AS reaches another network entity in the same AS via intermediate hops. To reach entities outside the AS, the inter-domain routing protocol or EGP (Exterior Gateway Protocol) used today is the Border Gateway Protocol or BGP [3]. Each AS announces aggregate information for the entities in its network via BGP to neighboring ASes. This is in the form of a routing announcement or routing update for one or more network prefixes. A network prefix is a representation of a set of IP addresses, such as $128.32.0.0/16$ for every address in the range of $128.32.0.0$ to $128.32.255.255$. Through the path vector operation of BGP, other ASes find out how to reach these addresses.

A packet that is sent from an AS X to an IP address in a different AS Z will traverse a series of links determined by multiple routing protocols. Firstly, the IGP inside AS X will determine how to send the packet to the nearest border router. The border router inside AS X will determine the inter-AS path via BGP, such as "AS X, AS Y, AS Z". The packet will then be sent to AS Y. AS Y will use BGP to determine that the next AS is AS Z. AS Y will use its IGP to send the packet across its network to the appropriate border router to send it to AS Z. AS Z will then use its IGP to send it to the destination inside its network.

Network traffic engineering tasks are critical to the operation of individual ASes. These tasks tune an operational network for performance optimization, and include traffic load balancing, link provisioning and implementing link fail-over strategies. For example, load balancing typically minimizes over-utilization of capacity on some links when other capacity is available in the network. In order to effectively traffic engineer a network, a traffic matrix is required. A traffic matrix represents the volume of traffic that flows between all pairs of sources and destinations inside an AS. However, due to a variety of reasons including limited network software and hardware capabilities, detailed network traffic information is often unavailable to build a traffic matrix. Thus a variety of techniques have been developed [4, 5, 6, 7] to *estimate* the traffic matrix from more easily obtainable network link load measurements. However, variations in BGP routes have the potential to add significant variability to the traffic matrix, which the prior work has not considered.

It has been approximately 15 years since BGP was deployed on the Internet. The number of ASes participating in BGP has grown to over $14,000$ today. This growth has been super-linear during the past few years [8]. With this sudden growth there has been concern in the research community about how well BGP is scaling. In particular, it has been noted that there is significant growth in the volume of BGP route announcements (or route flapping) [9] and in the number of BGP route entries in the routers of various ASes [8]. This has the potential to significantly impact packet forwarding in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGMETRICS/Performance'04, June 12–16, 2004, New York, NY, USA.
Copyright 2004 ACM 1-58113-664-1/04/0006 ...$5.00.

the Internet.

If the inter-domain path for reaching a particular destination keeps changing, then packets will traverse a different set of ASes after each change. Further, for an intermediate AS that peers with multiple ASes at different border routers in its network, changes in the inter-domain path will cause packets to traverse different paths inside its network to different border routers. This has several implications for the intermediate AS. Packet delivery times or latency within that AS can vary since the paths inside its network keep changing. Latency sensitive applications such as voice-over-IP can be adversely affected. If the intra-domain paths vary, then the traffic demands for different links in the network will vary. This variability in turn will impact the traffic matrix and make it's estimation more difficult.

In this paper, we answer the question "Do BGP routing table changes affect how traffic traverses a large IP network?". We study a "tier-1" ISP that connects to over $2,000$ other ASes. A significant percentage of Internet traffic transits this network. For these reasons, we believe that it is a suitable point for studying the impact of BGP on traffic inside an AS. We examine BGP data from multiple routers in the network. We correlate this with packet traces collected on several different days at different locations inside the AS. The contributions of our work are:

- We develop a methodology for analyzing the impact of BGP route announcements on traffic inside an AS. It separates inherent traffic dynamics such as time-of-day effects from egress PoP shifts due to BGP routing changes.

- We present results from the correlation of captured packets from an operational network with iBGP data. We find that a significant number of routing changes continuously occur. However, for the links that we measure, we experimentally conclude that these changes do not significantly impact the paths of most packets. Prior work [10] has found that only a small number of BGP announcements affect most of the traffic. However, even a few BGP changes can potentially significantly impact most of the traffic. We address what impact these few BGP announcements have.

The paper is organized as follows. We begin with related work in Section 2 followed by Section 3 that explains the problem we address in this work. We explain our methodology for tackling this problem in Section 4. We describe the data used in Section 5 and present our results in Section 6. In Section 7, we analyze the routing data and packet traces further to justify our findings. We end with conclusions in Section 8.

2. RELATED WORK

Due to the difficulty in collecting detailed data for all traffic in a large network, statistical inference techniques have been developed [4, 5, 6, 7] to obtain traffic matrices. These techniques attempt to infer the byte counts for origin-destination pairs within a network based on link byte counts. The traffic matrix that is estimated is one where the origins and destinations are routers inside the local network. In reality, for ISP networks, most of the origins and destinations are end hosts outside the local network. Thus inter-domain route changes between the end hosts can change the origin and destination routers inside the local network. This has the potential to reduce the accuracy of these techniques and thereby impact the traffic engineering tasks based on the estimated traffic matrices. Zhang et al. [7] identify this problem but assume it to be negligible based on their experience in the proposed generalized gravity model. We correlate BGP data with traffic measurements to quantify this effect.

Much of the prior work in inter-domain routing has been in analyzing aggregate statistics of eBGP (external BGP) tables and updates. To our knowledge, little prior work has focused on iBGP (internal BGP) behavior. Also, we study iBGP dynamics on the packet forwarding path in an operational "Tier-1" ISP, instead of prior work that studied related issues through simulations or controlled experiments. We are aware of only two studies [11, 10] that have correlated traffic measurements with BGP data from an operational network.

Uhlig and Bonaventure [11] use six successive days of traffic measurements and a single snapshot of a BGP routing table to study the distribution and stability of traffic. They find that traffic is not evenly distributed across ASes in terms of hop distance from the measurement point. They show that under 10% of ASes sent about 90% of the traffic. The largest ASes in terms of traffic contribution remained the largest from day to day.

Rexford et al. [10] associate the number of BGP updates with traffic behavior in a large "tier-1" network. They find that a small number of prefixes receive most of the BGP updates and that most traffic travels to a small number of prefixes. They find that the prefixes that carry most of the traffic do not receive many BGP updates. These results might lead one to conjecture that BGP routing updates do not cause significant traffic shifts. However, even if the prefixes that carry most of the traffic receive few BGP updates, these few updates can still cause significant egress border router changes. These results do not specifically demonstrate the extent to which BGP updates cause shifts in intra-domain traffic because the number of updates itself is not sufficient to understand this issue. Every BGP announcement can potentially change the attribute that determines the egress border router. Thus the number of BGP updates does not directly translate into the amount of traffic that shifts. In our work, we develop an entirely different methodology than used by Rexford et al. [10]. We perform a thorough study of how BGP updates can affect the intra-domain traffic matrix. We go beyond counting the number of BGP messages associated with popular prefixes to actually accounting for how every packet is affected by every BGP change. We measure the impact in terms of traffic variability in backbone links and quantify volumes of traffic shifts. We find that for some traffic, a few BGP updates do change the egress router address and cause the traffic to shift between intra-domain paths. However, most of the traffic is unaffected. The traffic we measure contains large flows that receive BGP updates carrying fewer egress router changes than those for other flows, which was not explored in the prior work.

3. PROBLEM DESCRIPTION

BGP is a path vector routing protocol that exchanges routes for IP address ranges or prefixes. Each route announcement has various components, such as the list of prefixes being withdrawn, or the prefix being added, the AS path to be followed in reaching the prefix, and the address of the next router along the path. Every AS that receives a route announcement will first apply its import policies [3] and then BGP "best" route selection, which takes into consideration preferences local to the AS, the AS path length, and the best IGP path to the border router, among others. If the route is selected, then it has the potential to be passed onto other neighboring ASes. Export rules or policies determine which AS may receive this announcement. The current AS will be added to the AS path and the next hop router will be changed to one of this AS's border routers.

Many ASes connect via BGP to multiple upstream ASes or ISPs,

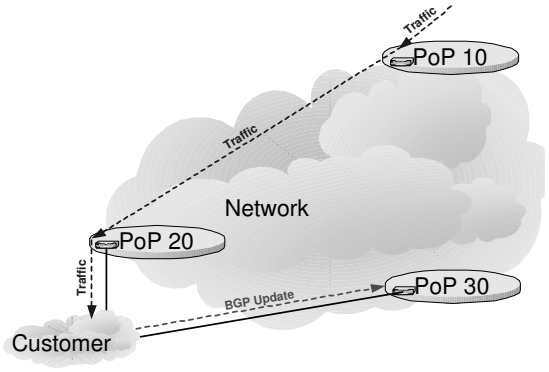

Figure 1: Intra-domain route and traffic through the network

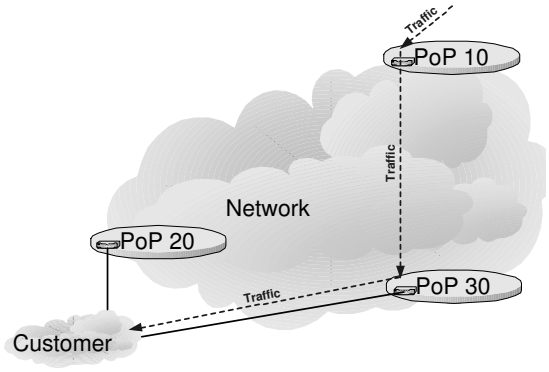

Figure 2: Intra-domain route and traffic through the network after BGP change

and even at multiple points to the same ISP. This trend, known as multihoming, has become very popular over the past few years, as indicated by the tremendous growth in BGP participation [12]. As a result, an intermediate AS may receive multiple routes in BGP for the same destination address prefix. This may cause the intermediate AS to keep changing the route it uses to reach this destination. This can occur due to many reasons. Each AS along a path applies local policies in accepting some routes and not others. BGP route selection is used to pick the "best" of the remaining routes via 13 steps [13]. In fact, each AS may have multiple BGP routers connected via internal BGP or iBGP [3], and different parts of the AS may be using different routes to reach the same destination. The concatenation of such route policy and selection rules across multiple routers in each of multiple ASes along a particular AS path to a destination leads to a very complex routing system [14]. Any portion of this system can contribute to route changes when faced with multiple choices to reach a destination. Rapid changes can make routing for a destination prefix unstable [15]. In addition, rapid changes can also significantly impact traffic patterns *within* an AS.

The "tier-1" ISP network that we study connects to multiple other ASes, in multiple geographically distinct locations called PoPs or Points of Presence (also known as switching centers). Such an ISP has a large and complex network of routers and links to interconnect these PoPs. Further, each PoP is a collection of routers and links that provide connectivity to customer ASes or peer ASes in a large metropolitan area. Routers within and across PoPs use iBGP to distribute BGP routes. iBGP is typically used in networks with multiple routers that connect to multiple ASes. It may not be possible to distribute the BGP routing table in IGP in a scalable fashion to all routers within large ASes [3]. Thus iBGP is used to exchange BGP routes among these routers and IGP is used to exchange routes for local addresses within the AS. An AS network may be designed under several constraints such as the average latency or jitter inside the network. Thus, the ISP will have to "engineer" its network to ensure that loss and delay guarantees are met. The task of traffic engineering may include setting IS-IS or OSPF link weights so that traffic travels along the shortest paths in the AS's network and congestion is limited. Over time, the traffic exchanged with these neighboring ASes may change. As a result, the link weights will have to be updated. Furthermore, the traffic exchanged with these ASes may grow and more customer ASes may connect to the ISP. As a result, more links will have to be "provisioned" into the network. These tasks are usually performed by first generating a "traffic matrix" which shows the traffic demands from any point in the network to any other point. It can be created at different levels - each row or column of the matrix can be a PoP or AS or router or ingress/egress link. PoP-to-PoP traffic matrices are important for provisioning and traffic engineering inter-PoP links, which typically require the most critical engineering.

If the inter-domain BGP route for a destination prefix changes, then the path of traffic to one of these destination hosts through the network may change. Consider the example in Figure 1 where traffic destined to the customer AS enters the network through PoP 10. The current BGP announcement from the customer determines that the "egress" or exit PoP for this traffic is PoP 20. Each BGP announcement has a next hop attribute that indicates the egress BGP router that traffic to the destination address can be sent to. Thus the announcement from the customer would indicate that the next hop is the egress BGP router in PoP 20. If a new BGP announcement is heard that changes the next hop router to one in PoP 30, then this traffic will travel to PoP 30 instead, as shown in Figure 2. As a result, the traffic is now traveling between PoPs 10 and 30 instead of PoPs 10 and 20. The path taken by this traffic inside the network may now be very different. The links between PoPs 10 and 20 will have less load and the links between PoPs 10 and 30 will have more. Congestion may occur and future growth of the network may be impacted. Further, due to this change, this traffic will now experience a different latency because it traverses a different path. Latency sensitive applications such as voice-over-IP may be adversely affected if such changes occur often.

If this happens frequently, estimating traffic matrices for this network may be more challenging than previously assumed. If flows between end hosts keep changing the origin and destination points inside the local network, then the byte counts between these points will keep changing. Without traffic matrices that can account for and represent such variability, traffic engineering will become harder. There is significant potential for such changes to occur. Of the over $2,100$ ASes that connect directly to the network, over $1,600$ have additional indirect paths via other ASes to reach the network. In general, over half of the non-ISP ASes on the Internet have multiple paths to the "tier-1" ISPs of the Internet [12].

Note that in this work, we only address the impact on traffic in relation to path changes inside the network. Some of these changes may also be associated with path changes inside other ASes and in the inter-AS path. This may result in changes to the congestion or packet delay experienced by traffic, which may even cause congestion control reactions or end user behavior to change. We account for these effects due to real routing changes in our methodology

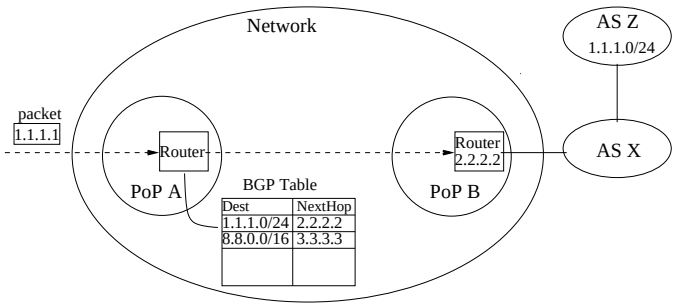

Figure 3: Data packet and BGP correlation example

by collecting and using real backbone traffic. However, we are unable to account for how the end user behavior would have been had there been no routing changes. Also, the problem we solve is only relevant in a typical network where most links are neither fully utilized nor empty, but have "moderate" utilization. If all the links are fully utilized, any shift in traffic flows will cause TCP algorithms to return to fully utilizing the link capacities, resulting in no change to the traffic matrix. Alternatively, if all the links have no load, then no traffic engineering tasks are needed and no traffic matrices need to be calculated.

4. ANALYSIS METHODOLOGY

4.1 Ingress PoP to Egress PoP Traffic Matrix

Since we wish to determine the impact of routing changes for traffic engineering and network capacity planning, we are only concerned with inter-PoP variations in traffic. Typically, each PoP is housed within a single building, and is a collection of routers and links between them. It tends to have a two-level hierarchical routing structure. At the lower level, customer links are connected to access routers. These access routers are in turn connected to a number of backbone routers. The backbone routers provide connectivity to other PoPs as well as other large ISPs. Installing additional capacity *within* a PoP (between access routers and backbone routers in the same PoP) is relatively less expensive and requires less planning and time compared to capacity upgrades between PoPs (between backbone routers in different PoPs). Thus we believe that intra-PoP links are rarely congested and intra-PoP variations are unlikely to cause significant latency variation.

If we are only concerned with variations in the inter-PoP paths that traffic takes across the network, we need to consider both traffic information and routing information at the granularity of PoPs. For a particular packet, an ingress PoP is the PoP where the packet enters the network, while the egress PoP is the PoP where the packet leaves the network, presumably toward the destination address. We need to determine if the egress PoP for any packet changes due to BGP route changes. Thus, we need to construct a PoP-to-PoP traffic matrix. Each column of the matrix corresponds to an egress PoP and each row corresponds to an ingress PoP. An entry in this matrix indicates how much of the traffic entering the corresponding ingress PoP exits the network at the corresponding egress PoP. Changes over time in this kind of traffic matrix indicates changes in traffic patterns between PoPs while ignoring changes in traffic patterns between links inside any PoP.

To generate this matrix, we need BGP routing information and packet headers. For every packet, we need to determine which PoP it will exit the network from. The destination address in the packet header indicates where the packet should finally go to. The BGP routing table entry for this destination address gives the last hop router inside the network that will send the packet to a neighboring AS. We use router address allocations and routing information specific to the network to determine the PoP that every egress router belongs to. In this fashion, we can determine the egress PoP for every packet. For example, consider Figure 3. At time t, a packet with destination address $1.1.1.1$ enters the network at PoP A. We use the BGP table from the ingress router in this ingress PoP to find the destination address $1.1.1.1$. This table indicates that the routing prefix is $1.1.1.0/24$ and the next hop router is $2.2.2.2$. This means that router $2.2.2.2$ inside the network will deliver this packet to a neighboring AS it and will eventually reach the destination prefix $1.1.1.0/24$. Using our knowledge of router locations and routing information specific to the network, we determine that $2.2.2.2$ is in PoP B. Thus we add the size in bytes of this packet to the (A, B) entry in the traffic matrix for time t.

4.2 Variability due to BGP

For traffic engineering and capacity provisioning, the traffic matrix needs to be considered. If this matrix varies a lot, it becomes harder to calculate it accurately and appropriately engineer the network. As has been observed in much prior work, Internet traffic has inherent variability, due to end-user behavior, congestion control and other reasons. However, there can be even more variability due to BGP routing changes. We want to identify the variability due to BGP, not the inherent traffic variability. By carefully using fresh versus stale routing data to calculate the traffic matrices, we can identify the variability that is due to BGP routing changes.

In the first scenario, we attempt to accurately account for what happens in the network. We maintain the latest BGP table for every point in time for a router by applying the BGP updates as they are received at the router. We call this the *dynamic* BGP table. For every packet that is received, we check this BGP routing table to find the egress PoP for that destination and update the traffic matrix. In this way, we can calculate the actual time-varying traffic matrix for the network that accounts for the combined effect of inherent traffic variability and changes due to BGP announcements.

In the second scenario, we consider what would happen if BGP changes did not occur. Here, we use a BGP routing table that existed at a previous point in time. We use this same *static* BGP routing table to calculate the traffic matrix for every point in time during the traffic measurements. This traffic matrix only accounts for the inherent traffic variability. We call this the "stale" traffic matrix.

We then subtract these two time-varying traffic matrices to obtain the changes to the traffic matrix that were only due to BGP announcements. We are only comparing the traffic at the same points in time between the actual traffic matrix and the "stale" traffic matrix. After subtracting the two matrices at some time t, we get the "difference" matrix for time t. Suppose that a cell at (A, C) in the difference matrix has value z. This means that at t, an extra z bytes from PoP A egressed at PoP C due to one or more previous BGP routing changes. There should be a corresponding $-z$ bytes for some other cell in the A row.

This can occur in the following scenario as in Figures 1 and 2. Suppose that at the start of our study, the egress PoP for the destination prefix $1.1.1.0/24$ was PoP 20. Suppose m bytes of packets travel to this destination prefix at time $t-2$, and at time $t-1$ a routing change occurs changing the egress PoP to PoP 30. At time t, z bytes of packets travel to this destination prefix. The "stale" traffic matrix will show $(10, 20) = m, (10, 30) = 0$ at time $t-2$ and $(10, 20) = z, (10, 30) = 0$ at time t. The traffic matrix with

routing changes will show $(10, 20) = m, (10, 30) = 0$ at time $t - 2$ and $(10, 20) = 0, (10, 30) = z$ at time t. The "difference" matrix will show $(10, 20) = 0, (10, 30) = 0$ at time $t - 2$ and $(10, 20) = -z, (10, 30) = z$ at time t.

Note that here we are only concerned with intra-AS changes due to BGP - i.e., shifts in the egress PoP within the network. BGP changes may cause inter-domain paths to change. The difference matrix removes the impact of inter-domain changes on traffic and only focuses on the impact due to intra-domain changes.

5. ANALYSIS DATA

We now describe the packet and BGP routing data that we collect from the network to understand if BGP routing changes impact how traffic traverses the network.

5.1 Packet Trace Collection

To build an accurate PoP-to-PoP traffic matrix for any significant amount of time, we need a tremendous amount of data. The network that we study has over 40 PoPs worldwide, and we need to create approximately a $40X40$ matrix. Some PoPs have hundreds of ingress links. Thus we would need to capture packet headers from thousands of ingress links. This is currently infeasible, due to multiple reasons including collection logistics, storage limits and computation time limits. Instead, we capture packet traces from multiple ingress links for several hours at different times, as shown in Table 1. We analyze our problem for each packet trace individually. Thus instead of building PoP-to-PoP traffic matrices, we build an ingress link to egress PoP vector for each packet trace, which we refer to as a traffic fanout. The sum of all the traffic fanouts from all the ingress links in a PoP forms a row of the traffic matrix. If each of the traffic fanouts is not affected by BGP changes, then the traffic matrix is unaffected, which makes it easier to engineer the network.

We capture packet traces using passive monitoring infrastructure. We use optical splitters to tap into selected links and collection systems that store the first 44 bytes of every packet. Every packet is also timestamped using a GPS clock signal, which provides accurate and fine-grained timing information. We pick multiple ingress links as shown in Table 1 in an attempt to obtain packet traces representative of the traffic entering the network from a single ingress PoP. The categorization of the neighboring ASes into "tiers" is based on the classification from Subramanian et al. [16]. The information in this table and the results that we present in later sections have been anonymized. The traces cannot be made publicly available to preserve the privacy of the network's customers and peers.

5.2 Approximations

A significant amount of computation time is required for the analysis of these traces. For example, trace D in Table 1 represents over 2.5 billion packets and consumes $162GB$ of storage. In order to keep computation times low, we employ one simplification technique and two approximations.

In the first approximation, instead of calculating and storing a separate traffic fanout for every instant in time during a trace, we create one fanout for every 20 minute period. That is, we aggregate all the packets received in every 20 minute window and calculate the traffic fanout due to those packets. The simplification technique here is that we do not treat packets individually, but rather treat them as a flow aggregate. For every 20 minute window, we group packets by the destination address, and lookup the egress PoP for this destination address once. This simplification avoids the overhead of looking up the same address multiple times when present in multiple packets with no loss in accuracy. In calculating the traffic matrix with routing changes, we use a BGP table snapshot at the start of every 20 minute window. We calculate a table snapshot by batching the routing table changes in a 20 minute window. We then use it to compute the egress PoP for traffic in the next 20 minutes. We then calculate a new table snapshot for the following 20 minutes of traffic and so on. Thus there may be some out-of-date routing information from one window to the next.

While we choose 20 minutes arbitrarily, we have experimented with smaller values down to 2 minute intervals. We find that this window size introduces negligible errors in the traffic fan-out calculation while smaller values significantly slow down the computation. For example, we randomly picked a 60 minute segment of trace D and analyzed the variability in 2 minute intervals. We saw no additional variability by volume than what the 20 minute analysis showed. However, we were not able to run the 2 minute analysis for the whole trace due to computation time.

The second approximation is that we only consider 99% of the traffic. More specifically, we only consider the largest *flow*s (packets grouped by the destination address) that account for at least 99% of the traffic. We have observed the phenomenon that there are a few flows that account for the majority of traffic and many flows that contribute an insignificant amount of traffic, as has been shown in prior work [17]. By ignoring the smallest flows that contribute a total of at most 1% of the traffic in any 20 minute window, we significantly reduce the computation overhead. For example, in trace D, only $30,000$ out of $200,000$ destination addresses carry 99% of the traffic. Thus, in each 20 minute window, we only lookup $30,000$ addresses in the routing table, instead of almost 10 times as many. Therefore this approximation makes the fan-out computation significantly faster at the cost of ignoring only 1% of the total traffic.

5.3 BGP Routing Collection

To determine which egress PoP a packet is sent to, we need to correlate the packet headers with BGP routing information. We collect BGP data from PoPs 8 and 10. We use the GNU Zebra [1] routing software to connect to each of these PoPs and collect routing updates. In the case of PoP 8, we connect to the same router that we collect packet traces from. The Zebra listener connects as an iBGP route reflector client and stores all route updates that are received. For PoP 10, the Zebra listener connects as a customer AS in an eBGP session. Each of the updates is timestamped to allow correlation with the packet traces that we collect. Each update corresponds to an actual change in the BGP routing table at the respective router. Thus we capture *all* the BGP routing changes that occur for the given router.

While we present data from the eBGP listener for comparison, we primarily focus on our iBGP data. iBGP data is richer than eBGP data in many aspects. It reflects both changes in BGP routes learned from external ASes by the network, and changes to BGP routes for internal addresses. It identifies the egress router within the network for any destination prefix, while an eBGP routing table from a particular collection router would only indicate the address of that collection router for all destination prefixes. iBGP data reflects changes in IGP routing as well, because if the IGP routing metric changes resulting in a change to the best next hop BGP router for a prefix, it will be seen as a change to the corresponding iBGP table entry, which would not be true of eBGP. Also, it includes some private BGP community attributes that help us determine the source of the routing announcements within the network,

[1]GNU Zebra Routing Software, http://www.zebra.org/

Table 1: Packet Traces

Trace	PoP	Link	Link Speed	Link Type	Neighbor	Date	Duration (hours)
A	8	2	OC-12	ingress	Tier-2 ISP	06 Aug 2002	6.1
B	8	2	OC-12	ingress	Tier-2 ISP	06 Aug 2002	9.9
C	8	3	OC-12	ingress	Tier-3 ISP	06 Aug 2002	6.4
D	8	1	OC-12	ingress	Tier-2 ISP	06 Aug 2002	22.4
E	8	3	OC-12	ingress	Tier-3 ISP	07 Aug 2002	9.6

which would not be seen in eBGP data.

6. RESULTS

While we have analyzed all the traces in Table 1, we will focus on the results from packet trace D for brevity. Our analysis for all the traces produced very similar results. We present trace D here since it is the longest trace.

6.1 Stability of the BGP Routing Table

We begin by considering how stable the BGP routing table is. If the routing table does not change at all, then it can have no negative impact on traffic within the network. In Figure 4, we show the number of BGP routing table changes for a typical week in PoPs 8 and 10. There were $765,776$ eBGP routing table changes and $1,344,375$ iBGP routing table changes during this week. Each point in the graphs shows the number of routing table changes during a 20 minute window. We see that the typical number of iBGP routing table changes is about 133 per minute, while eBGP changes occur at about half that rate. We observe that occasional spikes are interspersed among this continuous BGP "noise" of 133 changes per minute. During the spikes, the average number of iBGP routing changes is much higher, up to $6,500$ per minute.

In Figure 5, we show a histogram of the number of iBGP route changes during a typical week. We aggregate route changes into 20 minute windows. We plot the percentage of number of changes in each window on the vertical axis, with the horizontal axis showing the actual number of changes. In the bottom graph, the range of the horizontal axis is limited to $10,000$ in order to avoid distorting the shape of the graph with outliers. This figure illustrates the noise characteristic of route changes more clearly. The number of 20 minute intervals during which $1,000$ or fewer changes occurred is negligibly small. On the other hand there are $1,000 - 4,000$ changes per 20 minute interval for a majority of the entire week. Figure 6 plots the histogram of changes over a typical month. The shape is similar to that in Figure 5 which confirms that the distribution of route changes is similar on longer time-scales. We have verified this behavior over a period of several months.

The presence of continuous BGP routing table changes indicates that the Internet's routing infrastructure undergoes continuous change. Prior work has shown the amount of variability in eBGP, however little prior work has focussed on iBGP behavior inside an AS. This continuous change may be related to the size, complexity and distributed control of the Internet. Thus BGP updates have the potential to affect intra-domain traffic continuously, and not just during short periods of instability in the Internet. These short periods of update spikes are relatively infrequent, but we observe that they can cause a ten-fold increase in the rate of routing change. It is difficult to accurately identify the cause of such spikes. However, significant events such as router crashes, BGP session restoration and maintenance activities are likely causes. If an unusual event such as the loss of connectivity to a PoP or a major neighboring AS occurs, then significant traffic shift will naturally occur. In this

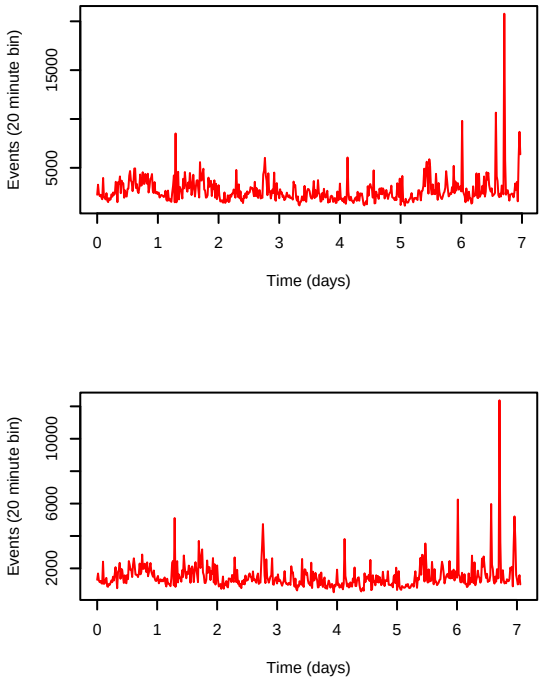

Figure 4: BGP routing table changes from Tuesday 06 August 2002 to Tuesday 13 August 2002 (iBGP on top, eBGP on bottom)

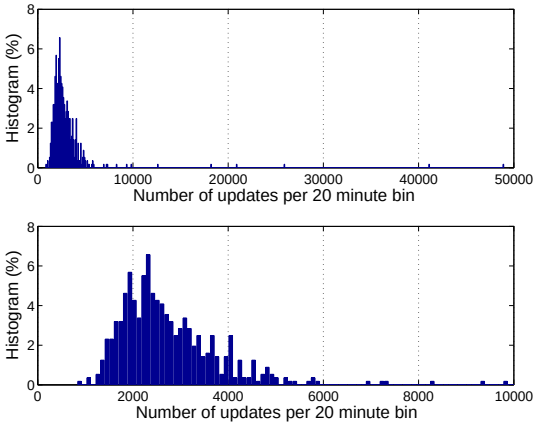

Figure 5: Histogram of iBGP route changes over a typical week

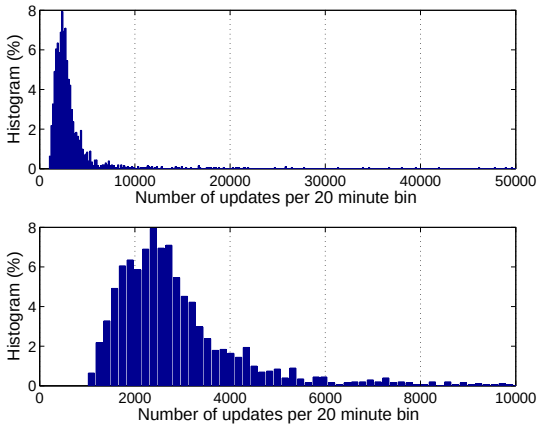

Figure 6: Histogram of iBGP route changes over a typical month

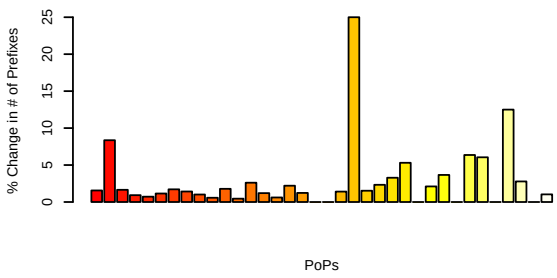

Figure 7: Changes in prefixes at each egress PoP during trace D

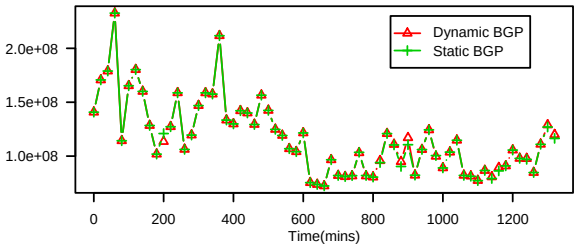

Figure 8: Bytes from trace D to PoP 2 for dynamic BGP table and static BGP table

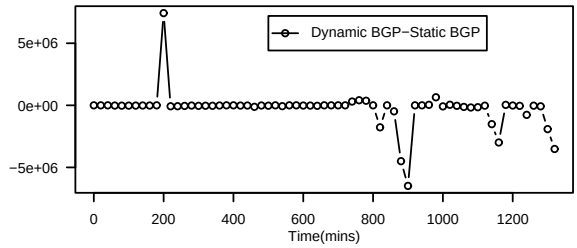

Figure 9: Difference in bytes from trace D to PoP 2 (dynamic BGP - static BGP)

work, we do not focus on these rare events but instead study the impact of routing table changes during more typical time periods. We confirm that no major loss of connectivity occurred during trace D by presenting Figure 7. We track the number of destination prefixes that exit each egress PoP. In this figure, we plot the maximum percentage change in this number for each PoP throughout the duration of the trace. We see that in most cases, less than 10% of the total prefixes exiting at each PoP were added or removed from the BGP routing table. This is typical behavior during the other traces that we analyzed. The two cases of 25% and 12.5% change were due to maintenance at two new egress PoPs being provisioned into the network. No traffic exited those two PoPs from trace D.

6.2 Overall Impact on Intra-Domain Traffic

We now investigate if this continuous noise of BGP routing table changes affects how traffic is forwarded in the network. Figure 8 shows the traffic volume per 20 minutes for packet trace D toward a particular egress PoP in the network. One line indicates the traffic computed with a static BGP table while the other is that with a dynamic BGP table. The fluctuations observed in both cases arise due to the variability inherent in traffic, such as due to user behavior. The difference between the two lines shows how much of this traffic shifted inside the network due to BGP changes. Since the two lines are very close to each other, this variability is negligible. Figure 9 plots the difference in the number of bytes toward the egress PoP for the two cases, by subtracting the value for the static BGP case from the value from the dynamic BGP case. The sum of this difference across all ingress links for each egress PoP forms

Table 2: Summary of Trace Results

Trace	# of Cells	Avg Shift per Cell	Std Dev of Shift	Cells With $> 5\%$ Shift	Volume Shift	Total Volume	% Volume Shift
A	648	0.17%	1.62	4	103 MB	398 GB	0.03%
B	1044	0.03%	0.24	0	58 MB	791 GB	0.01%
C	684	0.60%	7.53	4	33 MB	556 GB	0.01%
D	2412	0.07%	2.03	2	145 MB	1 TB	0.01%
E	1008	2.35%	15.05	24	144 MB	919 GB	0.02%

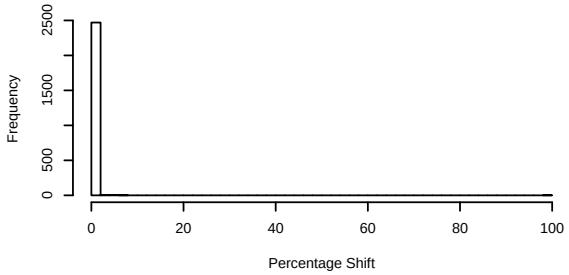

Figure 10: Histogram of egress PoP % traffic shift for trace D

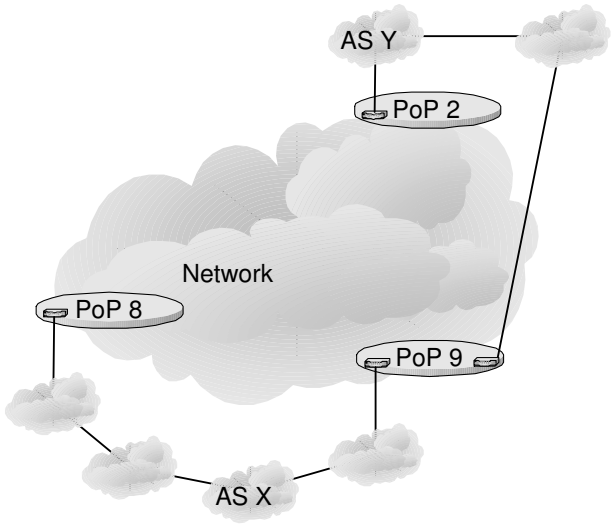

Figure 11: Traffic shift from PoPs 2 and 8 to PoP 9

the difference matrix that we previously described. We see that there is no difference for most of the time intervals. The maximum difference is about $7MB$ for any 20 minute window, compared to $120MB$ of traffic to this PoP at that time, which is only 5.8%. In Figure 10 we show a histogram of the number of time intervals across all PoPs for trace D by the percentage shift in traffic. We see that less than 5% of traffic shift occurred in almost all cases.

In Table 2 we summarize the results for all the traces. The second column shows the total number of cells or entries in the traffic fanout (i.e., the number of 20 minute time periods in the trace multiplied by the number of egress PoPs). The "Avg Shift per Cell" column shows the percentage of traffic shift averaged across all the cells and the next column shows the standard deviation of this value. The "Cells With $>$ 5% Shift" column shows how many of these cells had more than a 5% traffic shift. We find that the average shift over all time periods and PoPs is only 0.07% for trace D. In only 2 cases was the percentage shift more than 5%. However, in both cases, the actual volume of traffic that shifted was only several MB. From the last three columns in Table 2, we show that of the $1TB$ of traffic volume in trace D, only $145MB$ changed the egress PoP as a result of a BGP change, which is only 0.01%.

As shown by the last column, very small percentages of the ingress traffic move around due to BGP changes across all the traces that we analyzed. However, there are some cases where traffic from an ingress link to certain PoPs for certain time periods shifts. While these do not represent large volumes of traffic that can impact traffic engineering decisions, they can impact the performance of individual applications. Delay-sensitive applications such as voice-over-IP may experience degraded application quality due to traffic shifts between egress PoPs for individual prefixes. For example, a large volume of traffic toward a customer network $P1$ may shift frequently between two egress PoPs A and B, while the traffic toward another customer network $P2$ may shift in the reverse direction. While this may lead to very little change in the total volume of traffic toward egress PoPs A and B, customers $P1$ and $P2$ may experience significant delay fluctuations across the network. However we find that for our packet traces, the greatest number of shifts between egress PoPs across all flows (as defined in Section 5) is only 3. For example, in trace D, there were 67 20-minute windows, with an average of $23,409$ flows for 99% of the traffic in each window. An average of $5-6$ flows experienced a shift in the egress PoP per window. Therefore, only small numbers delay-sensitive flows are likely to experience fluctuations in quality across the network.

6.3 Specific Cases of Egress Shifts for Intra-Domain Traffic

We now examine two particular cases of variability in order to gain deeper insights into such occurrences. In trace D, about 42% of the total traffic variability involved only two destination networks. These two networks connect to the network we study in multiple places, as shown in Figure 11. This variability occurred between three PoPs that are spread across the east coast of the US. We found that traffic going to AS X shifted from the longer AS path via PoP 8 to the shorter AS path via PoP 9, while traffic to AS Y shifted from the shorter AS path via PoP 2 to the longer one via PoP 9. In each case, the BGP path changed only once throughout trace D. These changes in the inter-domain paths caused a change in the egress PoP for these destination addresses because different neighboring ASes peer with the network in different PoPs. In Figures 9, 12 and 13, we show the shift in traffic exiting at PoPs 2, 8 and 9. We can see that the dips in Figures 9 and 12 correspond to

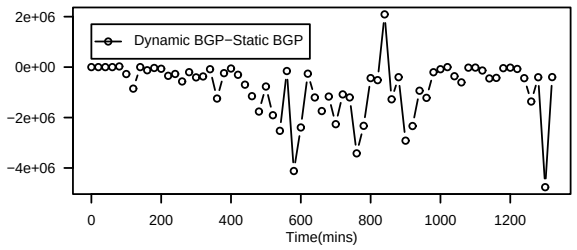

Figure 12: Difference in bytes from trace D to PoP 8 (dynamic BGP - static BGP)

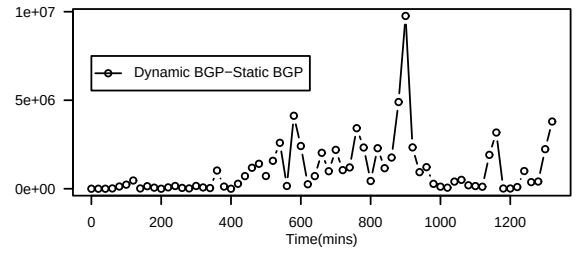

Figure 13: Difference in bytes from trace D to PoP 9 (dynamic BGP - static BGP)

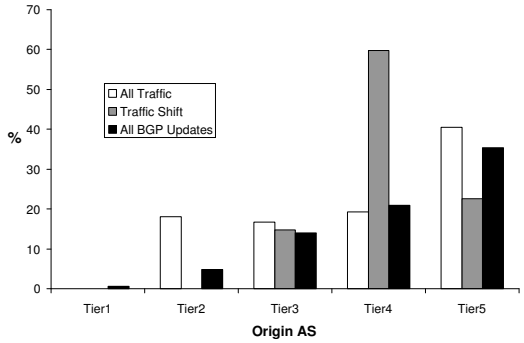

Figure 14: iBGP route changes, traffic and traffic shifts during trace D by origin AS

the peaks in Figure 13.

These two examples are typical of the variability in the fan-out to egress PoPs.

- We observe traffic shifting between different paths to multi-homed destination networks.
- Often the BGP path will change only once or twice during each trace.
- Only a few networks are involved in the majority of traffic shifts.

7. LIMITED IMPACT OF BGP CHANGES ON TRAFFIC

In the previous section, we showed that the traffic fan-out in the network is hardly affected by changes in BGP routes. Yet there is a significant amount of BGP activity all the time. In this section, we explain this discrepancy.

7.1 Distribution of BGP Changes and Traffic Across ASes

We begin by examining whether routing table changes, traffic and traffic shifts are similarly distributed across all the ASes. Since there are over $14,000$ ASes, we summarize the ASes into 5 distinct categories for simplicity. This categorization is based on Subramanian et al. [16]. Tier-1 ASes correspond to large global ISPs such as the one we study. Tier-2 ASes tend to be national ISPs, Tier-3 and Tier-4 are regional ISPs. Tier-5 ASes are stub networks that do not provide connectivity to other ASes. In general, a Tier-n AS is a customer of one or more Tier-(n-k) ASes.

In Figure 14, we compare BGP route changes, traffic destinations and traffic shifts for the origin ASes (i.e., the terminating AS along the path). We see that the majority of traffic is destined to Tier-5 ASes. This is consistent with the notion that the tiers provide connectivity to ASes except for Tier-5 stub ASes that house the end hosts. We see a similar trend with the number of BGP changes. Most of the routes that are affected are to prefixes terminating in Tier-5 ASes. However, we see that the traffic shifts are disproportionately more frequent for destination prefixes in Tier-4 ASes. This is due to a few networks being involved in the majority of traffic shifts, as we showed in the previous section.

In Figure 15, we compare the same distributions across the next ASes (i.e., the neighboring AS that traffic or paths go to). We see

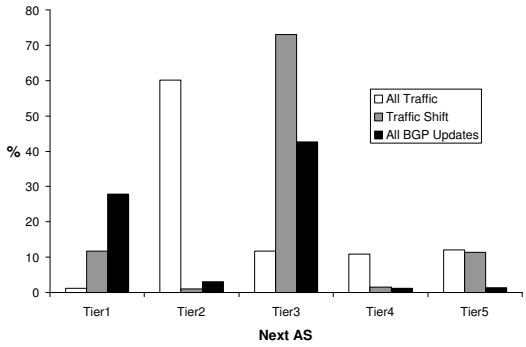

Figure 15: iBGP route changes, traffic and traffic shifts during trace D by next AS

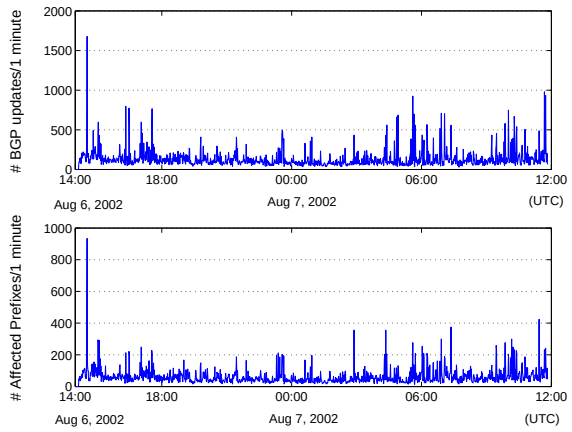

Figure 16: iBGP route changes and prefixes affected during trace D

that most traffic leaves the network to Tier-2 ASes. This is consistent with the notion that the network we study is a Tier-1 global ISP provides connectivity to many Tier-2 national ISPs. However, we see that the majority of BGP route changes are received from neighboring Tier-3 ASes. Consistently, the majority of traffic shifts involve neighboring Tier-3 ASes. Again, this is due to a few networks being involved in the majority of traffic shifts, as we showed in the previous section. Tier-1 ASes also account for a significant number of BGP changes. Since the network peers directly with Tier-1 ASes, and since these few ASes transit more prefixes than other ASes, tier-1 ASes show more BGP changes in Figure 15 than in Figure 14.

Thus we find that most traffic leaves the network to neighboring Tier-2 ASes and most traffic terminates at Tier-5 ASes. However, the traffic shifts are not distributed across these ASes in the same manner and the BGP changes are not distributed in the same way as traffic shifts. This can mean that either the BGP changes from each AS are not spread evenly across the BGP table or the BGP changes do not cause egress PoP changes. We now explore the first possibility and then explore the second possibility at the end of this section.

7.2 Distribution of BGP Changes Across the Routing Table

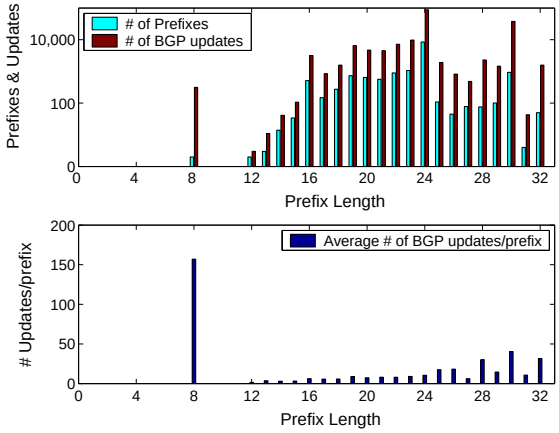

Figure 17: iBGP route changes and prefix length during trace D

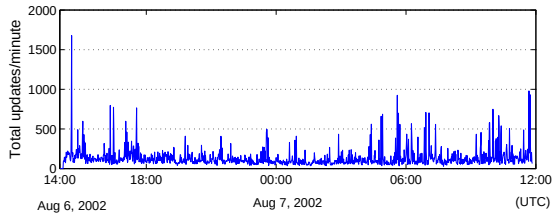

Figure 18: BGP route changes for all prefixes during trace D

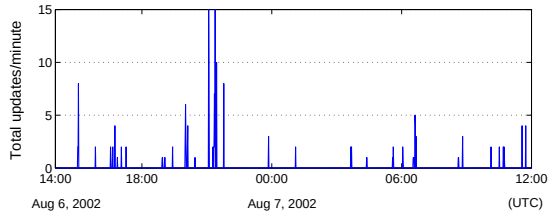

Figure 19: BGP route changes for heavy-hitters during trace D

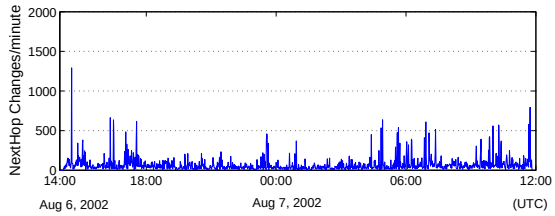

Figure 20: Next hop BGP route changes for all prefixes during trace D

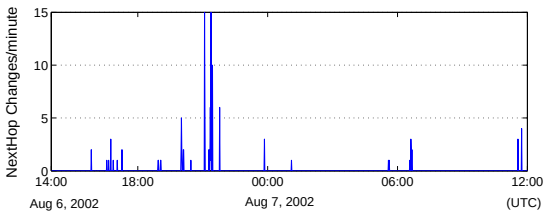

Figure 21: Next hop BGP route changes for heavy-hitters during trace D

In Figure 16, we show the number of routing table changes and the number of prefixes affected. We again see in the top graph that an average of 133 routing table changes occur every minute. In the second graph, we see that on average, roughly 70 routing table entries are affected every minute. Even during the spike of $1,500$ routing table changes early in the trace, only 900 routing table entries were affected. This shows that the same destination prefix can receive multiple routing changes within a short time period.

In Figure 17, we show the distribution of route changes with prefix length. From the top graph, we see that the number of changes (dark vertical bars) does not directly correspond to the number of routing table entries (light vertical bars) for each prefix range. In the second graph, we normalize the number of changes by the number of entries for each prefix. We see that $/8$ addresses receive an unusually high number of changes. $/8$ prefixes constitute less than 0.01% of the BGP table, but account for 18% of the route changes received. $/28$, $/30$, $/32$ address prefixes also receive a high number of updates per routing table entry. These more specific entries typically represent internal addresses within the network and customer networks that do not have a public AS number. They are usually represented in the eBGP routing table by a larger address range.

Thus we see that BGP routing table changes are not spread evenly across the routing table. Some routing table entries receive multiple changes, and entries of certain prefix lengths are more prone than others. Thus if most of the traffic is sunk by destination addresses that are in these change-prone prefixes, then there is more potential for shift.

7.3 Distribution of BGP Changes Across Traffic

Since BGP route changes are not spread uniformly across the routing table, and since subsequent traffic shifts are also not proportionately spread across neighboring and origin ASes, we now examine how BGP route changes are spread across traffic. Specifically, we examine which prefixes carry the majority of the traffic and examine how they are affected by BGP route changes. Prior work [18, 17] showed that network traffic contains heavy-hitters - i.e., a small set of destination network prefixes that together contribute a very large portion of traffic. We observed similar heavy-hitters in the packet traces we analyzed in this paper. In trace D, we found that $30,000$ addresses out of a total of $200,000$ in the trace accounted for 99% of the traffic, which is about 15% of the addresses. Only 1.5% of the addresses in the trace accounted for 80% of the traffic.

In Figure 18, we see again the number of iBGP route changes during trace D, with the average of about 133 changes per minute. In contrast, Figure 19 shows the number of changes for only the destination prefixes that account for at least 80% of the traffic. We see a significantly lower number of route changes. The maximum number of changes in any one minute interval is only 15, while across all prefixes, the maximum number is $1,600$. This shows that only a small fraction of the BGP route changes affect the majority of traffic. This is true of all the traces we examined in Table 1.

However, for our particular problem, we are only concerned with route changes that affect the next hop attribute. The next hop attribute determines the egress router, and thus the egress PoP, that traffic to a particular network prefix will go to. Only changes to this attribute can cause shift in the egress PoP for traffic. In Figure 20, we show the number of BGP route changes that affected the next hop for all prefixes. We see that the number of events has dropped to about half of that seen in Figure 18. Further, we are only concerned with changes to the next hop for the majority of traffic, which we show in Figure 21. Here we see an even smaller number of route changes that affect our problem of egress PoP shift. **Only 11% of the BGP changes for heavy-hitters caused next hop changes, while 63% of the BGP changes for all prefixes caused next hop changes**.

We conclude that heavy-hitters receive fewer route changes than most prefixes, and further, a significantly lower number of route changes for heavy-hitters causes next hop changes. For our problem, very few of the large number of route changes matter. **Only 0.05% of the total route changes during trace D caused next hop changes for heavy-hitter destination addresses**. These are the only ones that can potentially affect traffic fan-out toward egress PoPs, although in some cases the next-hop change may be from one router to another within the same egress PoP. This explains our findings that BGP route changes cause no more than 0.03% of traffic volume to shift the egress PoP.

There can be two reasons for this phenomenon. First, if a network prefix is unstable then packets traveling toward it may be frequently disrupted - during routing convergence, packets may be dropped, re-ordered or delayed. This can cause TCP sessions to back off and even terminate. Thus it could be that only stable network prefixes can sustain large, long traffic flows. Second, networks that attract large volumes of traffic may have more resources to afford good network administration and stable BGP configurations with their peers. Regardless of the cause of stability of heavy-hitters, there is a significant amount of instability for non-heavy-hitters. However, it is difficult to accurately determine the cause of the instability. Any of a large number of network events (from intra-domain IGP metric changes to router configuration changes in a neighboring AS) can cause a BGP change to occur. Since BGP is a path vector protocol, it is difficult to even determine the AS that originated a particular routing change, let alone the cause of it. Griffin [14] shows that a BGP network can nondeterministically change routing events in complex and non-intuitive ways as they are propagated. While it may be possible to study large numbers of correlated routing changes from multiple BGP vantage points, we believe it is difficult to accurately determine the cause behind the instability of individual destination prefixes.

8. CONCLUSIONS

Recent studies of BGP have shown a significant growth in the size and dynamics of BGP tables. This has led to concerns about what impact these trends in BGP have on the Internet. We focus on this issue for a large ISP. Large ISP networks are designed and maintained on the basis of metrics such as latency and the need to provision the network for future growth and changes in traffic. This engineering is typically based on calculating a traffic matrix to determine traffic demands for different parts of the network. Fluctuations in BGP routes can cause this traffic matrix to change, invalidating the engineering effort. Further, latency sensitive applications

can be adversely affected.

We have correlated iBGP route changes with packet traces in a large IP network to measure the variability in traffic fan-out from ingress links to egress PoPs. We have presented a methodology that separates the variability inherent in traffic from the variability that is due to BGP changes within an AS. From our analysis of several packet traces and associated BGP changes, our findings are:

- There is continuous iBGP noise of more than a hundred routing changes per minute. This is interspersed with rare periods of high changes, as much as several thousand per minute. eBGP changes occur at about half this rate.
- Hardly any fraction of the volume of traffic from any ingress link that we measured experienced an egress PoP shift due to BGP changes. At any time, only several flows experienced an egress PoP shift out of typically tens of thousands of flows in a trace. Affected flows experienced no more than a few shifts.
- Only few networks tend to be involved in a significant fraction of the traffic shift. This involves the inter-domain path changing, resulting in the intra-domain path changing.
- BGP route changes are not distributed evenly. Some route entries receive multiple changes, and some are more likely to receive a change than others. BGP changes and traffic seem similarly distributed by origin AS, while BGP changes and traffic shifts seem similarly distributed by next AS. Relatively few BGP changes affect the majority of the traffic, and even fewer change the egress PoP.

The traffic fanout is largely unaffected by BGP routing changes for several links that we have considered. If these links are representative of all the ingress links, then it is reasonable to assume that the traffic matrix is unaffected. Therefore it is possible to perform network engineering and provisioning tasks without concern for the effect of global routing changes. BGP changes are unlikely to cause latency variations within an AS for most traffic. Yet, some open issues remain. Are heavy-hitters relatively immune from change due to the engineering of networks or do TCP dynamics dictate that only stable routes can support heavy-hitters? Unstable routes may reflect connectivity that is undergoing rapid changes or heavy packet loss, and that may cause the TCP congestion control algorithm to cut back its sending rate. Another open issue is that since there is so much BGP change, the cause and origin of such updates should be understood. However, due to the non-link state nature of BGP, it is difficult to accurately identify the cause of individual updates. Correlation between successive updates from several locations in the Internet may provide better success. This work explored a specific effect of BGP changes inside a network, i.e. variability in intra-domain traffic patterns, during normal network behavior. Periodic network maintenance and link failures may cause additional impact on traffic that we have not explored.

9. ACKNOWLEDGEMENTS

Our work would not have been possible without the help of Sprintlink Engineering and Operations in collecting BGP and packet traces from the network. We also thank the anonymous reviewers for their detailed feedback.

10. REFERENCES

[1] D. Oran, "OSI IS-IS intra-domain routing protocol," RFC 1142, IETF, February 1990.

[2] J. Moy, "OSPF version 2," RFC 1583, IETF, March 1994.

[3] J. W. Stewart, *BGP4: Inter-Domain Routing in the Internet*. Addison-Wesley, 1998.

[4] Y. Vardi, "Network tomography: estimating source-destination traffic intensities from link data," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 365–377, 1996.

[5] C. Tebaldi and M. West, "Bayesian inference on network traffic using link count data," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 557–576, 1998.

[6] J. Cao, D. Davis, S. Wiel, and B. Yu, "Time-varying network tomography : Router link data," *Journal of the American Statistical Association*, vol. 95, pp. 1063–1075, 2000.

[7] Y. Zhang, M. Roughan, N. Duffield, and A. Greeberg, "Fast accurate computation of large-scale IP traffic matrices from link loads," in *Proc. ACM SIGMETRICS*, June 2003.

[8] G. Huston, "Analyzing the Internet's BGP Routing Table," Cisco Internet Protocol Journal, March 2001.

[9] T. Bu, L. Gao, and D. Towsley, "On routing table growth," in *Proc. IEEE Global Internet Symposium*, 2002.

[10] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP routing stability of popular destinations," in *Proc. Internet Measurement Workshop*, 2002.

[11] S. Uhlig and O. Bonaventure, "Implications of interdomain traffic characteristics on traffic engineering," *European Transactions on Telecommunications*, 2002.

[12] S. Agarwal, C.-N. Chuah, and R. H. Katz, "OPCA: Robust interdomain policy routing and traffic control," *Proc. IEEE International Conference on Open Architectures and Network Programming*, 2003.

[13] Cisco Systems Inc., "Cisco IOS IP and IP routing command reference, release 12.1."

[14] T. G. Griffin, "What is the sound of one route flapping?," in *Network Modeling and Simulation Summer Workshop, Dartmouth*, 2002.

[15] Labovitz, Malan, and Jahanian, "Internet routing instability," in *Proc. ACM SIGCOMM*, 1997.

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

[17] K.Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, and C. Diot, "A pragmatic definition of elephants in Internet backbone traffic," *Proc. Internet Measurement Workshop*, 2002.

[18] S. Bhattacharyya, C. Diot, J. Jetcheva, and N. Taft, "Pop-level and access-link-level traffic dynamics in a tier-1 PoP," in *Proc. Internet Measurement Workshop*, 2001.