BGP Routing Stability of Popular Destinations

Jennifer Rexford, Jia Wang, Zhen Xiao, and Yin Zhang
AT&T Labs–Research; Florham Park, NJ

Abstract— **The Border Gateway Protocol (BGP) plays a crucial role in the delivery of traffic in the Internet. Fluctuations in BGP routes cause degradation in user performance, increased processing load on routers, and changes in the distribution of traffic load over the network. Although earlier studies have raised concern that BGP routes change quite often, previous work has not considered whether these routing fluctuations affect a significant portion of the traffic. This paper shows that the small number of popular destinations responsible for the bulk of Internet traffic have remarkably stable BGP routes. The vast majority of BGP instability stems from a small number of unpopular destinations. We draw these conclusions from a joint analysis of BGP update messages and flow-level traffic measurements from AT&T's IP backbone. In addition, we analyze the routing stability of destination prefixes corresponding to the NetRating's list of popular Web sites using the update messages collected by the RouteViews and RIPE-NCC servers. Our results suggest that operators can engineer their networks under the assumption that the BGP advertisements associated with most of the traffic are reasonably stable.**

I. Introduction

An Autonomous System (AS) is a collection of routers and links operated by a single institution. Routing between ASes depends on the Border Gateway Protocol (BGP) [1], a path-vector protocol that operates at the level of address blocks, or prefixes. Each prefix consists of a 32-bit address and a mask length (e.g., $192.0.2.0/24$ consists of addresses from $192.0.2.0$ to $192.0.2.255$). Neighboring ASes use BGP to exchange update messages about how to reach different prefixes. A router can send an announcement of a new route for a prefix or a withdrawal of a route that is no longer available. Each advertisement includes the list of ASes on the path, along with several other attributes. Since routers do not send updates unless something has changed, the Internet could conceivably reach a "steady state" where no routers need to send new update messages. However, BGP routing in the Internet is far from stable.

BGP routing changes happen for a variety of reasons. The exchange of update messages depends on having an active BGP session between a pair of routers. Equipment failures or reconfiguration may trigger the closing of the BGP session, forcing each router to withdraw the routes learned from its neighbor; after reestablishing the session, the routers exchange their routing information again. Each router applies local policies to select the "best" route for each prefix and to decide whether to advertise this route to the neighbor. Changes in these policies can trigger new advertisements. A group of ASes may have conflicting policies that lead to repeated advertising and withdrawing of routes [2, 3]. In addition, intradomain routing or topology changes may cause some routers to select new BGP routes and advertise them to neighboring ASes.

BGP routing changes can cause performance problems. A single "event," such as a link failure, can trigger a long sequence of updates as the routers explore alternate paths. During this convergence period, the packets headed toward the destination prefix may be caught in forwarding loops. Exchanging and processing the update messages also consumes bandwidth and CPU resources on the BGP-speaking routers in the network. In addition, the new advertisements from neighboring ASes may change the paths that traffic takes through the network. This can cause congestion on certain links in the AS. Frequent changes in the advertisements from other domains make it difficult for operators to engineer the flow of traffic through an AS. For example, a BGP routing change may cause traffic to a particular destination prefix to leave the AS through a different egress point. If BGP routing changes affect a large portion of the traffic, past information about BGP updates would not be a sound basis for future operations decisions.

Early measurement work discovered an alarming number of unnecessary update messages, due to design choices made by router vendors [4, 5]. Changes in vendor implementations led to a significant decrease in the rate of BGP updates, though the number remained high. When routing changes occur, the BGP system can experience high convergence delay and a large number of update messages [6]. Delayed convergence is marked by high delay and high loss for IP packets to the affected destinations. Despite the large number of updates, a large fraction of destination prefixes have remarkably stable BGP routes [7]. A relatively small number of prefixes are responsible for the bulk of the BGP update messages. Several recent studies have made a similar observation about traffic volumes—a small fraction of the destination prefixes are responsible for the majority of Internet traffic [8–10].

This paper asks a simple, yet important, question about the relationship between BGP updates and traffic volumes: Do the small fraction of popular prefixes have relatively stable BGP routes? On the one hand, popular destinations would presumably have reliable and well-managed connections to the Internet, which would argue for greater stability. On the other hand, these destinations may have connections to multiple upstream providers, which may result in more BGP routes and the possibility of more frequent changes; in addition, operators in some ASes might intentionally change the routes for these popular destinations to engineer the flow of traffic. We also consider a second, closely related question: Is there a direct correlation between traffic volumes and BGP routing stability. That is, are prefixes receiving a higher volume of traffic more (or less) stable than prefixes receiving a lower volume of traffic? In short, we find that the answer to the first question is "yes" and the answer to the second question is "no."

The remainder of the paper is structured as follows. Section II analyzes BGP update data from AT&T's commercial IP backbone and the RouteViews and RIPE-NCC repositories, and shows that most of the update events are associated with a small fraction of the prefixes. Then, Section III joins AT&T's BGP data with traffic measurements from the AT&T backbone, and shows that most of these destination prefixes receive a very small fraction of the traffic. In Section IV, we generate a list of destination prefixes associated with the NetRating's top-25 Web sites and show that the BGP routes for popular prefixes are typically stable for days or weeks at a time. Section V concludes the paper with a discussion of future research directions.

II. BGP Routing Stability

This section describes the routing data used in our analysis. We define an update "event" as a collection of update messages for a single prefix that are spaced close together in time, from a single vantage point. We show that a few prefixes are responsible for the vast majority of events.

A. Collection of BGP Update Messages

Our study draws on BGP update messages from the publicly-available RouteViews [11] and RIPE NCC [12] servers and a BGP monitor in the AT&T backbone. The public servers collect update messages by establishing eBGP (exterior BGP) sessions with routers in participating ASes. As such, these logs provide a view of the "best" BGP-learned route as seen by these routers. The AT&T data are collected by a Zebra software router that has an iBGP (interior BGP) session with several BGP route reflectors. In our analysis, we focus on the data from a single route reflector; this data provides a view of the best route to every prefix from that vantage point. Our study uses update data from the entire month of March 2002.

The update data have a number of anomalies that can affect the analysis of routing stability. First, the BGP session that connects the monitor to an operational router may reset due to failures, reachability problems, or congestion. Reseting the monitor's BGP session results in a burst of update messages that do not necessarily reflect real routing changes. Second, some routers send redundant advertisements for the same prefix or withdrawals for prefixes that have not been advertised to the BGP neighbor. These reflect router implementation choices that trade off complexity and memory overhead for extra update messages. We preprocess the BGP data to remove the extraneous update messages since they do not reflect real BGP routing changes. For each dataset, we start with an initial BGP routing table and apply the stream of update messages to construct a view of the routing table at each point in time. We discard update messages that do not affect the contents of the table. This preprocessing step typically removed between 7% and 30% of the updates, though the number was higher for some of the RIPE sessions.

B. BGP Update Events

A simple count of the number of updates is not necessarily a good way to compare the routing stability of different prefixes. Upon receiving a new update message, a router may explore several alternate routes during the convergence process. The number of BGP update messages and the convergence delay may vary dramatically, depending on the timing of the messages and where the data are collected [6, 13]. Instead of counting messages, we identify updates that are spaced close together in time and collapse them into a single *event* associated with the destination prefix. We cannot be sure that our notion of an event actually corresponds to a single action, such as an equipment failure or routing policy change. Still, combining bursts of updates reduces the sensitivity of the results to the timing details and where the data were collected.

Our definition of an event requires a threshold for the time between updates for the same prefix. Figure 1 plots the cumulative distribution of the inter-arrival times of update messages. The RouteViews and RIPE curves each correspond to a single eBGP neighbor. Inter-arrival times of around 30 seconds are quite common for these two datasets, probably due to the min-route-advertisement

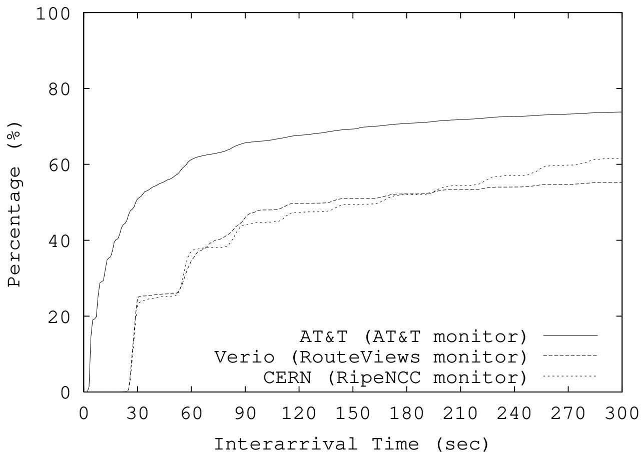


Fig. 1. Interarrival time of update messages for prefixes

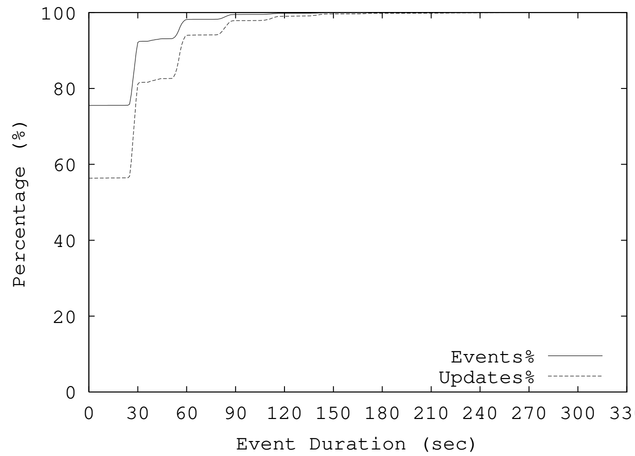


Fig. 2. Event duration for Verio session to RouteViews

timer that limits the rate of updates from an eBGP neighbor [6]; a popular vendor uses a default timer value of 30 seconds. Inter-arrival times of around 4 seconds are common for the AT&T data. This is likely due to the default min-route-advertisement timer for an iBGP neighbor. Our analysis of events considers a timeout threshold of 45 seconds; results for a 75-second timeout were very similar.

C. Analysis of BGP Update Events

The vast majority of events last for less than five minutes, as shown by the top curve in Figure 2. This curve plots the cumulative distribution of the duration of update events for the RouteViews data for a 45-second timeout. These results are consistent with previous analysis of BGP convergence delay [13], and with results for our other datasets. Although most events are short-lived, the small number of long events contain a larger number of update messages. The bottom curve plots the proportion of update messages that belong to events that are shorter than a certain duration. For the Verio session, no event lasted more than 328 seconds. In a few rare cases in other datasets, we see events of longer duration that could stem from flaky equipment that repeatedly goes up and down (within the 45-second timeout) or from routing policy conflicts that lead to persistent oscillation. As such, we consider both event frequency and duration as metrics in our analysis.

A small number of prefixes are responsible for most of the update events, as shown in Figure 3. The graph ranks the prefixes by the number of update events and plots the percent of events from the highest ranked prefix to the lowest. The three curves correspond to different timeout values for defining an event. The top curve has a zero-second timeout that counts each update message as a separate event; this curve is consistent with [7], which showed that a small fraction of prefixes contribute the bulk of the

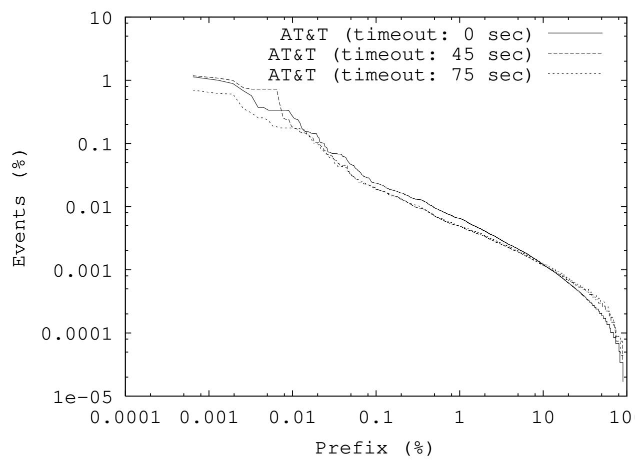


Fig. 3. Fraction of events vs. prefixes (AT&T data)

update messages. We also computed the total duration of all update events for each destination prefix. Plots of event duration (not shown) show the same trend as Figure 3. That is, we find that a small fraction of the prefixes have a much larger update duration than the remaining prefixes.

III. UPDATE EVENTS VS. TRAFFIC VOLUME

This section describes the traffic measurements from the AT&T network. We verify that a small fraction of prefixes receive the bulk of the traffic. Then, we combine the traffic and routing data to show that the vast majority of the events stem from a small number of unpopular prefixes, and that the popular prefixes do not experience many events.

A. Flow-Level Traffic Measurement

We identify the volume of traffic associated with each destination prefix from measurements of AT&T's peering links. The data were collected by enabling Cisco's Sampled Netflow feature [14] on the routers. The routers in each PoP (Point-of-Presence) were configured to send

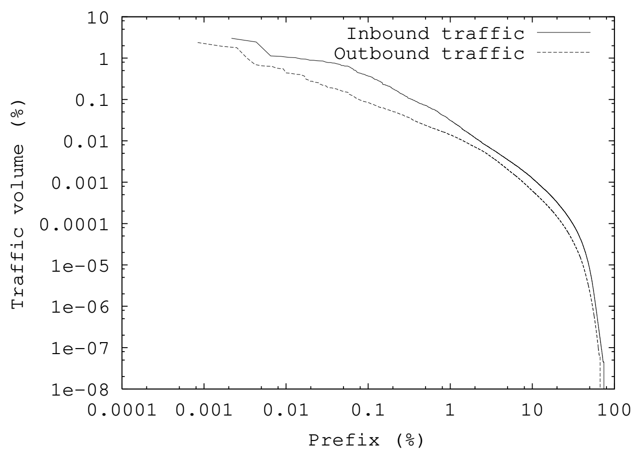


Fig. 4. Fraction of traffic vs. fraction of prefixes

the flow-level measurement records to a local collection server. Each server was configured to compute the hourly traffic volumes for each destination prefix. Separate traffic statistics were maintained for traffic entering and leaving the AT&T backbone. To reduce the processing overhead, each server applied the stratified sampling technique described in [15]. For our analysis, we combined the hourly data from March 2002 into a single traffic volume for each destination prefix in the inbound and outbound directions.

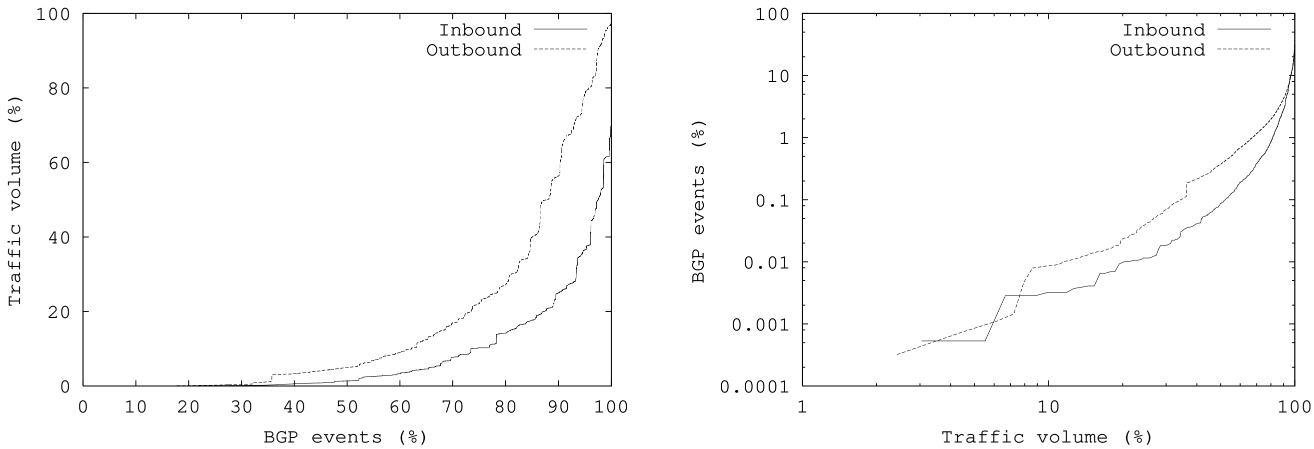
The outbound statistics represent traffic *sent* by AT&T customers to the rest of the Internet, whereas the inbound statistics represent traffic *received* by AT&T customers from the rest of the Internet. We classify the prefixes based on the community strings in the BGP advertisements; routes received from peers are tagged with a different community than those from customers. A small number of prefixes for multi-homed customers appear in both categories, since the "best" route varied over time with traffic occasionally exiting via the peering links. Each dataset provides a way to rate the popularity of different prefixes. For both sets of prefixes, the majority of the traffic travels to a small fraction of the destination prefixes, as shown in Figure 4; the rest of the prefixes received little or no traffic. The graph ranks the prefixes by the volume of traffic and plots the proportion of the traffic from the highest ranked prefix to the lowest. For example, the most popular prefix in each direction contributes more than 1% of the traffic, consistent with other studies [8–10].

B. BGP Stability and Prefix Popularity

The prefixes responsible for most of the events do not receive much of the traffic, as shown by Figure 5(a). To construct this graph, we ranked the prefixes by the number of update events (using a 45-second timeout) and plotted the cumulative distribution of the volume of traffic destined to the prefixes associated with these events. The curves grow steeply to $y = 100\%$ at the end because some of the traffic is associated with prefixes that have few events, if any. For the inbound curve, half of the update events are associated with prefixes that receive 1.4% of the traffic. This disparity exists for two reasons. First, these update events stem from a relatively small fraction of the prefixes, consistent with Figure 3; 50% of the events are associated with just 4.5% of the prefixes. Second, these prefixes do not receive much traffic. If all prefixes were equally popular, then these prefixes would contribute 4.5% of the traffic; instead, they contribute only 1.4%.

There are two possible explanations for these results. First, unstable BGP routes make it difficult for other hosts to reach these destinations. Even if the destinations are reachable during an event, routing changes can cause transient forwarding loops that result in packet delay and loss. Still, this is not the entire explanation. Upon further inspection, these prefixes were not receiving much traffic, in general, even when the BGP routes appeared to be stable. These prefixes are not especially popular and may have fairly unreliable (or poorly-managed) connections to the rest of the Internet. Although we can say that the unstable prefixes tend to be unpopular, the reverse is not necessarily true. In general, we do *not* see a direct correlation between traffic volume and BGP routing stability. Many of the low-volume prefixes had very few events and did not have any long-lived events. In fact, some of these prefixes may be statically injected into BGP by the service provider, rather than managed individually by the owners.

The popular prefixes responsible for most of the traffic do not experience many events, as shown by Figure 5(b). We constructed this graph by ranking the prefixes by the volume of traffic and plotting the cumulative distribution of the events contributed by these prefixes. The curves rise so slowly that we have displayed this graph as a log-log plot. For the inbound curve, 50% of the traffic traveled to destination prefixes that contribute only 0.1% of the events; this traffic is associated with 0.25% of the prefixes. Similar results hold for event duration (not shown). One explanation for these results is that the popular destinations have reliable, well-managed equipment and that problems, when they arise, are detected and fixed quickly.



(a) Traffic volume vs. update events (b) Update events vs. traffic volume (log-log plot)

Fig. 5. Relationship between traffic popularity and routing stability (CDF)

IV. Popular Web Sites

In this section, we construct a set of popular prefixes using NetRating's list of top-25 Web sites. We show that update events for these prefixes (and for the popular prefixes from the Netflow data) are infrequent and short-lived.

A. Identifying Prefixes for Popular Web Sites

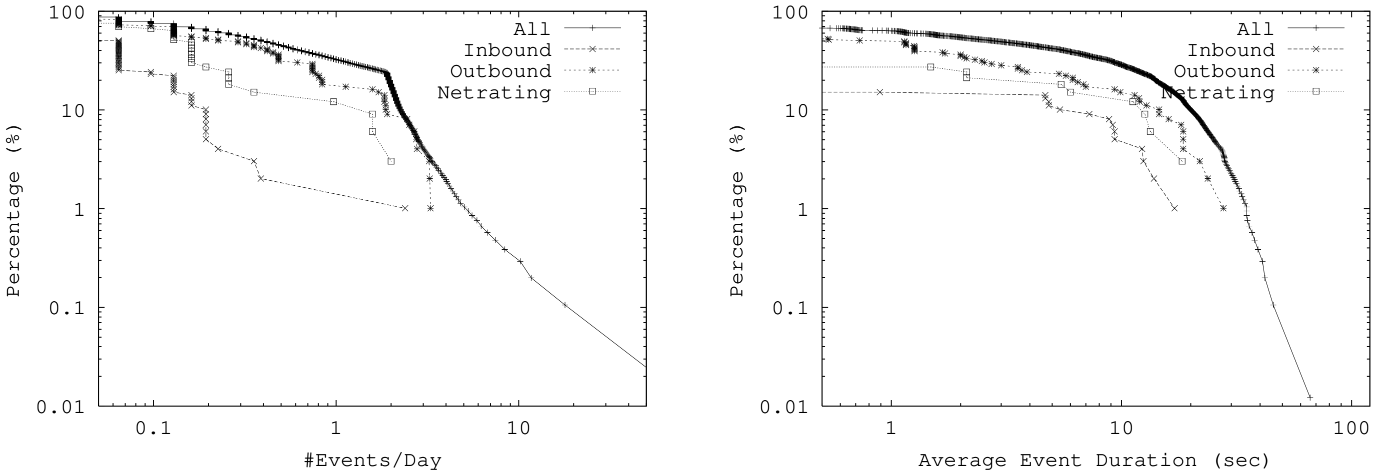
One limitation of the analysis in Section III is that the high-volume prefixes on AT&T's peering links may not be universally popular. To complement our results, we constructed a set of popular prefixes using the publicly-available NetRatings rankings. First, we took the NetRatings list of the top-25 Web sites [16] for "at home" users from the end of 2001. Second, we generated a list of domain names, typically by adding "www" and "com" to the site name (e.g., converting "Amazon" to "www.amazon.com"). This resulted in a list of 25 domain names for Web sites that send (and presumably receive) a large amount of traffic, and represent institutions with a large commercial interest in having high availability.

Third, we generated a set of IP addresses associated with each site by querying DNS servers all over the world. We started with a list of more than 50,000 DNS servers from the work in [17]. These DNS servers come from a wide range of locations in the Internet, and each DNS server has an IP address associated with a different destination prefix. We randomly generated a smaller list of 5% of these DNS servers. Fourth, we used "dig" to send queries to these DNS servers to translate each of the domain names into an IP address. Fifth, we identified the longest matching prefix in the BGP data for each address. This resulted in a final list of 33 prefixes.

B. BGP Stability for Popular Web Sites

Figure 6(a) plots the complementary cumulative distribution of the average number of events/day for all prefixes, the top 100 from the inbound Netflow data, the top 100 from the outbound Netflow data, and the 33 from NetRatings. The graph is plotted on log-log scale to emphasize the tail of the distribution. The bottom three curves show that *none* of the "inbound", "outbound", and "netrating" prefixes had more than 3 events per day, on average. The vast majority of these prefixes had fewer than 0.2 events per day, resulting in five or more days between successive events, on average. In contrast, more than 1% of the "all" prefixes had more than five events per day.

In addition to having fewer events, the popular prefixes tended to have *shorter* events, as shown in Figure 6(b). In fact, 85% of the "inbound" prefixes had an average event duration of *zero*; these prefixes had events with a single update, or no events at all. This number was 73% for the "netrating" prefixes, 47% for the "outbound" prefixes, and only 32% for the "all" prefixes. None of the "inbound", "netrating", and "outbound" prefixes had an average event duration of more than 20 seconds, suggesting that most events stemmed from BGP routing convergence rather than long-term oscillation. However, 0.1% of the "all" prefixes had an average event duration exceeding 40 seconds. We saw similar results for the other BGP sessions to the RouteViews and RIPE monitors. We speculate that most events for the "inbound", "outbound", and "netrating" prefixes involve changing from one route to another, whereas some "all" prefixes sometimes become unreachable, requiring the routers to explore and ultimately withdraw several alternate routes during convergence.



(a) Average rate of update events

(b) Average duration of update events

Fig. 6. Update events for Verio (AS 2914) session to the RouteViews server (CCDF with log-log plot)

V. Conclusion

Despite the large number of BGP update messages, popular prefixes tend to have stable BGP routes for days or weeks at a time. The vast majority of the update events are concentrated in a few prefixes that do not receive much traffic. These results have important traffic engineering implications—operators can assume that the BGP routes corresponding to most of the traffic are reasonably stable.

In ongoing work, we are investigating why the "inbound" prefixes have consistently fewer (and shorter) events than the "outbound" prefixes, and why the NetRatings prefixes tend to fall in between these two curves. We also plan to study how BGP stability and prefix popularity vary over time, and over different time-scales. Finally, we plan to conduct traceroute experiments to understand how (and whether) forwarding instability relates to BGP instability.

Acknowledgments

We would like to thank Tim Griffin for the AT&T BGP data, the software for parsing BGP updates, and cautioning us to preprocess the update data. We also thank Carsten Lund for the aggregated Netflow data, Oliver Spatscheck for the list of DNS servers, and Glenn Fowler for the software for computing the longest prefix match. We are also grateful to the RouteViews and RIPE NCC projects for public access to BGP update data. Thanks also to Matt Grossglauser for his comments on a draft of the paper.

References

[1] Y. Rekhter and T. Li, "A Border Gateway Protocol." Request for Comments 1771, March 1995.

[2] K. Varadhan, R. Govindan, and D. Estrin, "Persistent route oscillations in inter-domain routing," Tech. Rep. 96-631, USC/ISI, February 1996.

[3] T. G. Griffin and G. Wilfong, "An analysis of BGP convergence properties," in *Proc. ACM SIGCOMM*, September 1999.

[4] C. Labovitz, R. Malan, and F. Jahanian, "Internet routing stability," *IEEE/ACM Trans. Networking*, pp. 515–528, October 1998.

[5] C. Labovitz, R. Malan, and F. Jahanian, "Origins of pathological Internet routing instability," in *Proc. IEEE INFOCOM*, 1999.

[6] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," *IEEE/ACM Trans. Networking*, vol. 9, pp. 293–306, June 2001.

[7] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of Internet stability and wide-area network failures," in *Proc. Fault-Tolerant Computing Symposium*, June 1999.

[8] W. Fang and L. Peterson, "Inter-AS traffic patterns and their implications," in *Proc. IEEE Global Internet*, December 1999.

[9] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: Methodology and experience," *IEEE/ACM Trans. Networking*, vol. 9, June 2001.

[10] N. Taft, S. Bhattacharyya, J. Jetcheva, and C. Diot, "Understanding traffic dynamics at a backbone POP," in *Proc. Scalability and Traffic Control in IP Networks, SPIE ITCOM*, August 2001.

[11] "Route Views Project." `http://www.routeviews.org`.

[12] "RIPE NCC RIS." `http://www.ripe.net/ripencc/pub-services/np/ris-index.html`.

[13] C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja, "The impact of Internet policy and topology on delayed routing convergence," in *Proc. IEEE INFOCOM*, April 2001.

[14] "Sampled Netflow." `http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120limit/120s/120s11/12s_sanf.htm`.

[15] N. Duffield, C. Lund, and M. Thorup, "Charging from sampled network usage," in *Proc. Internet Measurement Workshop*, November 2001.

[16] "NetRatings." `http://www.netratings.com`.

[17] Z. Mao, C. Cranor, F. Douglis, M. Rabinovich, O. Spatscheck, and J. Wang, "A precise and efficient evaluation of the proximity between Web clients and their local DNS servers," in *Proc. USENIX*, June 2002.