Understanding BGP Next-hop Diversity

Jaeyoung Choi, Jong Han Park, Pei-chun Cheng, Dorian Kim, Lixia Zhang

University of California, Los Angeles
Computer Science Department
Technical Report # 100026
Aug. 12, 2010

Understanding BGP Next-hop Diversity

Jaeyoung Choi*, Jong Han Park†, Pei-chun Cheng†, Dorian Kim‡, Lixia Zhang†

*Seoul National University, jychoi@mmlab.snu.ac.kr

†University of California, Los Angeles, {jpark,pccheng,lixia}@cs.ucla.edu

‡NTT Communications Inc., dorian@blackrose.org

***Abstract*— BGP is the de facto routing protocol of the global Internet to exchange reachability information between routers and autonomous systems. For each destination, a BGP router selects and propagates only a *single* best path to its neighbors. Conceptually, each BGP router may learn only one or a few paths for a given destination even when a much larger number of alternative paths exist, which leads to a common concern that the lack of alternative paths can reduce a network's robustness to failures as well as flexibility in traffic engineering, and can lead to slow adaptation to topological changes. However there has been no quantitative measurement to assess the degree of path diversity, or lack of it, in the operational Internet. In this paper we use iBGP routing data collected over two years from a Tier1 ISP, ISP_A, to quantify BGP next-hop diversity for all destinations. Our results show that ISP_A can reach the majority of prefixes through multiple next-hop routers: 88% of prefixes can be reached via more than 2 next-hop routers and 78% of prefixes could be reached via more than 5 next-hop routers. Through several case studies of prefixes with different diversity degrees, we identify two major factors that impact the number of observed next-hop diversity: the Tier1's local route preference and the number of peering points between large ISPs. Moreover, we observed that a small fraction of prefixes have a very high degree of next-hop diversity (>=30) which is due to specific topological connectivity conditions. Although our results are derived from BGP data collected in a single ISP, it sheds lights on the major factors that impact path diversity and can serve as valuable input into the ongoing IETF efforts to increase BGP path diversity such as Best-External, Add-Path, and Diverse-BGP-Path.**

***Index Terms*—BGP; Path Diversity; Tier1 ISP; Measurement**

I. Introduction

Border Gateway Protocol (BGP) [18] is the routing protocol used in the Internet today. Although a BGP router may learn multiple paths from its peers for a given destination, BGP specification requires the router to select and propagate only one single best path. As a consequence, the amount of alternative paths that a BGP router can learn is limited, making the network less robust to failures and less flexible in load balancing and traffic engineering. Furthermore, this incomplete view of the network has unanticipated negative side effects such as persistent route oscillations [11] in the worst case.

BGP is known to suffer from slow convergence [15], and together with this limited visibility of alternative paths, the amount of packet losses can increase during the delayed convergence time to explore alternative paths that were invisible when the best path fails [24]. The packet losses can then translate into a degraded data plane performance, especially for real-time applications such as video streaming and VoIP [14, 19]. Recently, addressing this problem of slow BGP convergence is receiving much more attention, as the number of real-time applications that demand a higher quality of service have increased.

Today, a rapid advancement in hardware and software technology have made it feasible for BGP routers to support more than the default "one-destination-one-path" design, and there are active discussions and proposals on increasing BGP path diversity to improve robustness and performance [17, 9, 16] in Internet Engineering Task Force (IETF).

However, while the research and operation community put forth avid efforts in increasing the number of BGP paths, there has been little understanding and measurement efforts on quantifying the *existing* path diversity, or more specifically, the *lack* of path diversity in the operation networks. Previous works focused on measuring the AS level path diversity (*i.e.* , the distinct number of AS level paths towards a given prefix, the disjointness of paths) by simulations based on a small set of prefixes [22].

In this work, we measure an operation oriented diversity metric, *next-hop diversity*, as observed from all the *backbone routers* in a Tier1 ISP (ISP_A) for *all prefixes in the global routing table*. In this paper, we make the following contributions:

- We define and quantify next-hop diversity using iBGP data collected inside ISP_A, we find that, without any modifications to BGP, more than 84% of all prefixes can be reached via more than 2 peering locations (*i.e.* PoP - point of presence), and the majority of prefixes can be reached via multiple next-hop routers: more than 88% and 78% of all prefixes can be reached via more than 2 and 5 next-hop routers respectively.
- Through case studies, we analyze the prefixes with different level of diversities, and find that next-hop diversity is determined by the local routing preference and number of peering points between ISP_A and its neighboring ASes in general. Furthermore, we find that the prefixes with very high next-hop diversity are mostly caused specifically by the lack of geographical presence of ISP_A in some regions. This observation suggests that, to assure high diversity for all prefixes effectively without wasting

valuable router resources, it is necessary to modify BGP such that it can selectively increase the diversity on focused prefixes, rather than simply adding more BGP paths to all prefixes.
- Based on the historical data collected over two years, we study the evolution of next-hop diversity in time. We find that most of prefixes change their next-hop diversity in a seemingly unpredictable manner, which is a collective result of inter-domain connectivity and routing decisions. However, we observed that the maximal next-hop diversity slowly increases in time. Further investigation verifies that this observation is mainly due to the increased number of backbone routers in the Tier1 ISP.

This paper is organized as follows. Section II provides a brief review on BGP that is particularly relevant to our study. Section III describes our data set, definition of next-hop diversity, and how we computed next-hop diversity. We start by presenting our results of measuring next-hop diversity at a time instance in Section IV. Then in Section V, we present out results on next-hop diversity at different times and study how next-hop diversity has changed in time. In Section VI, we compare our work with previous work. In Section VII, we discuss about implications of our results, and finally in Section VIII, we summarize and conclude our paper.

II. Background

In this section, we provide a brief overview on BGP operations that are particularly relevant to our study, followed by the description of path diversity as used in this paper. Then, we describe iBGP hidden path phenomenon and explain how it hides alternative paths and reduces the number of overall visible paths. Interested readers are referred to [18, 12] for more detailed descriptions on the design principles of BGP and its operations.

A. Routing in the Internet

The Internet is made of tens of thousands of different networks called Autonomous Systems (ASes) and BGP is the glue that connects them together. Routers in different ASes set up BGP sessions in between to exchange routing information (inter-domain routing). Such sessions are called eBGP (external BGP) sessions. BGP sessions are also used to disseminate BGP routing updates within the same AS (intra-domain routing), and these sessions are called iBGP (internal BGP) sessions.

All routing protocols have means to prevent routing loops. In eBGP, routers detect routing loops at inter-AS level by inspecting the AS_PATH attribute carried in BGP messages. A router will drop a BGP message if AS_PATH already contains its own AS number. To avoid routing loops in iBGP, iBGP requires that all BGP routers within the same AS be directly connected to each other to create a full mesh, and that reachability information learned from any iBGP speaker must not be forwarded to any other iBGP speaker.

The following example shows how eBGP works together with iBGP. In Figure 1, AS2 maintains eBGP sessions with AS1, AS3, and AS4. The number of eBGP sessions between ASes depends on how much redundancy the ASes want to have. For example, there is only one eBGP session between AS2-AS3 (over R_{24}-R_{31}) and AS2-AS4 (over R_{23}-R_{41}), whilst there are two eBGP sessions between AS1-AS2 (over R_{11}-R_{22} and R_{12}-R_{21}). Within AS2, a fully-meshed iBGP configuration is used to distribute the reachability information.

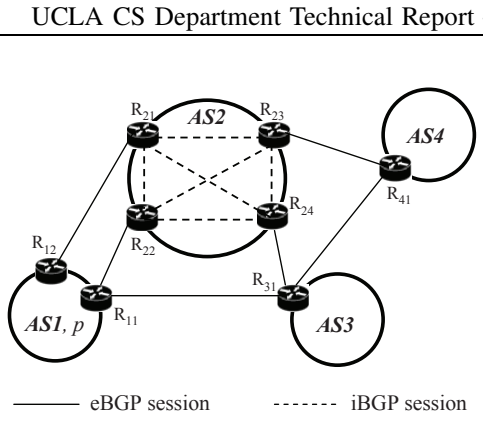


Fig. 1. BGP Operations in the Internet

When AS1 announces the reachability of its prefix p over the eBGP sessions to AS2 and AS3, AS2 will distribute this reachability information to all iBGP routers within the network (namely R_{21}, R_{22} ,R_{23}, and R_{24}) using the full-mesh iBGP sessions. After receiving this reachability information, R_{23} and R_{24} will further propagate this information to AS4 and AS3 respectively so that these neighbor ASes can also reach prefix p. This process repeats in every AS throughout the Internet until all ASes learn how to reach prefix p announced by AS1.

B. iBGP architectures

The original design of iBGP required all member routers connect in a full mesh, and this led to scalability problem in the network provisioning due to the non-linearly increasing number of iBGP sessions per added router. To mitigate this scalability problem, two schemes have been proposed: AS confederation [21] and Route Reflection [7]. We briefly describe each one of them below.

AS Confederations: AS confederation architecture groups a number of routers together into subASes. This leads to many subASes within a domain, and each subAS communicates with each other within the domain just as in eBGP. Within each subAS, iBGP speakers must be fully meshed. Routing loops are avoided between ASes via a new BGP attribute called AS_CONFED_SEQ, which works similarly to the AS_PATH attribute in eBGP.

Route Reflection: Route reflection architecture consists of a route reflector server (RR) and route reflector clients (RR clients). Under a route reflection architecture, non-RR iBGP routers connect to a route reflector server. The non-RR routers send updates to their RR, and the RR will reflect this route to all of its other clients. To avoid forming a routing loop, route reflection defines new attributes, namely CLUSTER_LIST and ORIGINATOR_ID, and use them in the similar way that

AS_PATH and AS numbers are used in eBGP.

Depending on the iBGP architecture deployed, the number of paths visible to reach a destination can differ. In both AS confederations and RR, only the best paths are further propagated from one side of subASes and route reflectors boundary to another side. This reduction in the visible number of paths leads to the reduced path diversity. The full-mesh, although not scalable, preserves the optimal path diversity. It is worth mentioning here that quantification and analysis of path diversity under different architectures and the degree of differences would be helpful to clearly understand the trade-offs between the architectures, and the work is currently underway.

C. Path Diversities in BGP

A BGP message reveals path diversities at two different levels: AS and next-hop level, which we refer as AS-path diversity and next-hop diversity respectively.

AS-path diversity: As briefly mentioned in Section II-A, a BGP message carries AS_PATH attribute which records the AS level path through which the message traveled to reach the receiving AS. Each AS paths represents an AS level path to reach such destination. For example in Figure 1, the announcement of the reachability to prefix p in AS1 will arrive at AS4 both through AS2 and AS3. Thus, AS4 learns two different paths (AS4-AS2-AS1 and AS4-AS3-AS1) to reach prefix p. Retaining multiple AS paths in AS4 could be helpful in case of a failure occurring outside of AS4. For example, if AS2 fails, AS4 will still be able to forward the data packets destined to prefix p to AS3, which will in turn forward them to AS1. However, as the receiving end, an operator has little control on the number of visible AS paths to reach a given destination. The alternative AS paths for a given destination may be hidden by the neighboring ASes due to various reasons such as policy, and the distributed nature of BGP routing protocol does not allow an operator to have much influence on the AS paths that are not propagated by the neighboring ASes.

Next-hop diversity: BGP announcement messages for a given prefix can be received from multiple AS neighbors (*i.e.* next-hop AS), potentially leading to a high AS-path diversity. Furthermore, there can also be multiple routers (*i.e.* next-hop routers) to reach each of these neighboring ASes across different cities, which we refer as Point of Presence (*i.e.* next-hop PoP). For example in Figure 1, AS2 receives the reachability information on prefix p through both R_{21} and R_{22} from AS1, and BGP distinguishes these different paths to reach p in AS1 using an attribute named NEXT_HOP.

Maintaining visibility to multiple next-hop routers could be helpful in case of internal failures either on the paths to reach a particular next-hop router or the failure of the next-hop router itself. For example, when R_{12} fails in Figure 1, routers in AS2 can use R_{22}-R_{11} and will still be able to reach AS1. Between neighboring ASes, an operator is able to increase or reduce next-hop diversity. When higher next-hop diversity is desired, the operator could deploy more routers to peer with the neighboring ASes.

In general, a higher path diversity at both AS and next-hop level is desired for the purpose of robustness to internal and external failures, traffic engineering, and faster convergence. For example, when an AS (or a router) along the selected path fails and a BGP router has an alternative path in its routing table, the router can fail-over to the alternative path immediately without waiting for the convergence. In addition, a high degree of next-hop diversity offers operators flexibility to direct their traffic for better resource utilization (*i.e.* load balancing).

In this work, as a first step to understand the existing BGP path diversity, we limit our focus on next-hop diversity and analyze the factors affecting the amount of visible next-hop diversity.

D. Path Poisoning and Hidden Paths in iBGP

As mentioned in Section II-A, iBGP is originally proposed to connect all iBGP routers in a full-mesh to avoid routing loops and not to forward reachability information learned from another iBGP peers. However, this requirement leads to the "path poisoning phenomenon", in which an iBGP router withdraws all known, but non-best paths for a given destination.

For example in Figure 1, R_{24} learns three paths to reach p in AS1: through its direct links to AS1 (R_{12}-R_{21}-R_{24} and R_{11}-R_{22}-R_{24}) and through AS3 (R_{11}-R_{31}-R_{24}). Because BGP prefers the path with shortest AS_PATH length assuming the preference from the proceeding criteria are equal, R_{24} chooses the path announced by R_{22} (due to the shortest AS_PATH length and the shortest IGP distance within AS2) as its best path. R_{24} will withdraw this path from all iBGP speakers in AS2, since the path through AS3 is known but not selected as the best path. As a consequence, this path is known to R_{24} only. Even though AS2 uses full-mesh iBGP configuration which is known to preserve the optimal path visibility, iBGP routers are limited in obtaining the complete view of all feasible paths to reach a given destination due to this hidden path phenomenon. Recently proposed modification to BGP, external best feature [16] specifically addresses this issue. Although not all, this feature mitigates the problem of hidden path phenomenon to some degree.

E. Policy Routing

Route selection and propagation in eBGP are generally determined by networks' routing policies, in which the business relationship between two connected ASes plays a major role. AS relationships can be generally classified into customer-provider, peer-to-peer, or siblings.

In general for a given prefix, a route announced by customers is preferred over that announced by peers. The peer route, in turn, is preferred over a provider route. This preference is due to the economic incentives: when sending traffic over a customer or peer route, the sender is not charged

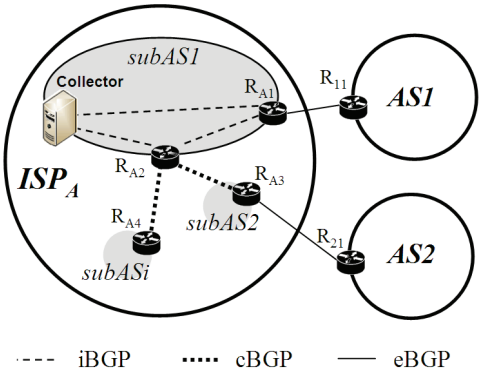


Fig. 2. High Level Topology of ISP_A

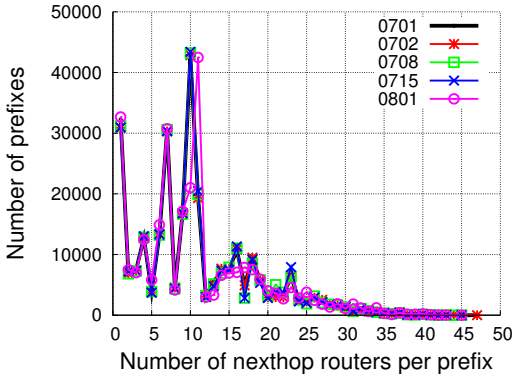


Fig. 3. Verifying Representativeness of Dataset

whereas the sender would be charged when using a provider route.

ISPs usually implement this policy using a BGP attribute named LOCAL_PREF. According to the best path selection rule, a route with higher LOCAL_PREF is preferred. Thus, an operator can implement this policy by configuring the routers such that LOCAL_PREF would honor this policy (*e.g.* the highest LOCAL_PREF value for a customer route).

This policy reduces the number of visible routes from inside an AS. As explained above in Section II-D, an iBGP router withdraws visible route that is not selected as the best path. Although a router sees a peer or provider route, it will hide the route until the selected best path fails.

III. Methodology

We used iBGP data collected from a Tier1 ISP (ISP_A) to quantify and analyze next-hop diversity inside its network. In this section, we first describe the high level network topology of this ISP. Then, we discuss the data collection settings and how we measure next-hop diversity.

A. A Brief Description of ISP_A's Topology

ISP_A is a Tier1 ISP in the Internet and has more than one hundred iBGP backbone routers. The routers are distributed globally across 14 countries in 3 different continents. To scale with the network size, ISP_A uses AS confederations [21]. All backbone routers belong to one specific subAS in their AS confederations configuration, and are connected in a full-mesh. Figure 2 depicts the topology of ISP_A at a high level, where $subAS_1$ represents the backbone network of this ISP.

In most cases ISP_A uses one of the backbone routers to set up BGP sessions with neighbor ASes. For example in Figure 2, R_{11} in AS1 establishes an eBGP session directly with R_{A1} in ISP_A and exchanges BGP messages. In some large PoPs, however, the routers are organized into several different subASes. In this case, prefixes are announced from the neighbor ASes to the backbone via at least one subAS. In Figure 2, R_{21} in AS2 establishes an eBGP session with R_{A3} in ISP_A and exchanges BGP messages. This message is further propagated to one of the backbone routers, R_{A2}.

An ISP often configures their routers not to propagate certain paths. ISP_A also applied such inbound filtering on routers at the boundaries of different continents. However, we verified with the operator that the number of such prefixes is relatively small and should not affect the generality of our measurement results.

B. Data Collection and Pre-processing

A collector (an iBGP router) is deployed in the backbone subAS as depicted in Figure 2, and the collector maintains iBGP peering sessions with all other routers in $subAS_1$ to passively record all iBGP updates received. The collected update messages and the snapshots of the routing tables are periodically stored to files in MRT [4] format every 15 minutes. We used *bgpparser* [1] to extract NEXT_HOP BGP attribute field to compute path diversity for a given prefix.

We exclude two types of prefixes from this measurement study: internal prefixes and potential bogon prefixes. Internal prefixes are meant to be used only inside ISP_A. Since the goal of our measurement is to understand the path diversity of commonly visible prefixes to all ASes in the Internet, we filter out such internal prefixes. In addition, we exclude the prefixes with its length greater than 24 because the BGP messages containing reachability information on these prefixes could have been filtered by BGP routers before reaching ISP_A, and can lead to inaccurate results.

C. Measuring Path Diversity

From the collected iBGP data, we gathered routing table snapshots (RIBs) from all iBGP peers. From each RIB entry, we extracted NEXT_HOP and AS_PATH attributes to calculate how many unique next-hop routers along with their geographical locations and next-hop ASes are visible to the collector for each destination. We note that ISP_A does not use *next-hop-self* option. Therefore, NEXT_HOP attribute contains the IP address of the router residing in the neighboring AS. Since all routers in the backbone are connected in a full-mesh, the number of visible nexthops observed by each backbone router should be the same to that observed by the collector.

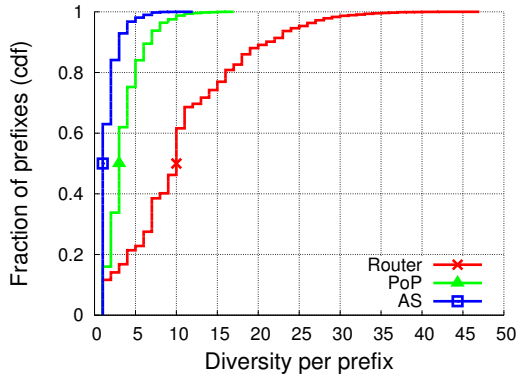


Fig. 4. Distribution of Next-hop Diversity

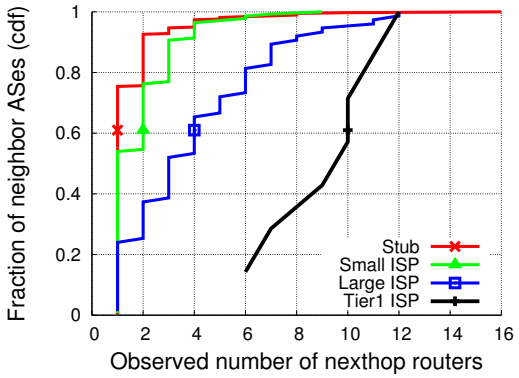


Fig. 5. Observed Connectivity of Different Neighbor Types

To study the next-hop diversity of a time instance, we chose to use the routing table snapshots taken from all backbone routers on July 1st, 2009. In addition, to ensure that these snapshots are representative, we also measured next-hop diversity using routing table snapshots taken at different times. Figure 3 compares the number of unique next-hop routers from routing table snapshots on July 1st, 2009 with those from four other snapshots on July 2nd (one day after), 8th (one week), 15th (two weeks), and August 1st, 2009 (one month) respectively. As depicted in this figure, the distribution of the number of next-hop routers for a given prefix from five snapshots are very similar. In addition, we checked that the total number of prefix entries in each snapshot and the set of unique neighboring ASes are roughly the same. Note that we performed the same measurements on all the dates shown in Figure 3, however due to space limit, in the following sections, we only present the results on July 1st.

IV. Next-hop Diversity

In this section, we start our work by quantifying the next-hop diversity for all prefixes of ISP_A with three different granularities: next-hop ASes, PoPs, and routers. Then we focus on characterizing and analyzing the router level diversity, which represents the essential unit of operational opportunities for failure recovery, traffic engineering, etc. In the following section, without further specification, we use *next-hop* and *next-hop diversity* to refer to the next-hop router and the router level diversity respectively.

A. Quantifying Next-Hop Diversity

1) Next-hop ASes: We first measure, for each prefix, how many next-hop (*i.e.* , neighboring) ASes can be used to reach a given network destination. Figure 4 shows the cumulative distribution (CDF) of the number of next-hop ASes to reach a prefix. For the total 276,712 prefixes, we observe that around 62% of all prefixes are reached via 1 neighbor, and almost all prefixes (about 96%) can be reached via less than or equal to 5 neighboring ASes.

Note that the number of next-hop ASes represent a gross diversity at the inter-domain routing level. For those prefixes that can only be reached through one neighbor, when this neighboring AS fails, ISP_A must wait for BGP to explore and settle down the routes via other neighbors (if there is any). The prolonged convergence delay in this case can potentially degrade the performance in the data plane [24]. However, such number of next-hop ASes only describe an abstract reachability at a high level. In operation, two ASes can peer with each other at different geographical locations using multiple BGP routers, which is the deciding factor for the real operational diversity.

2) Next-hop PoPs and Routers: In this section, we further measure the number of available next-hop routers and their geographical locations to reach a given destination as defined earlier in Section III.

Figure 4 shows the distribution of the number of observed next-hop PoPs and routers to reach each destination prefix. We observe that even though 18% of prefixes can still be reached via only one PoP from one neighboring AS, the majority of the prefixes can be reached via 2 to 5 PoPs. Furthermore, given that there exist multiple routers in a same PoP, that next-hop router diversity is further amplified and varies widely from 1 up to 47. Most of the prefixes (88%) have more than 2 next-hop routers, and around 47% of all prefixes have their next-hop router diversity between 6 and 11. There also exists a small fraction of prefixes (1.6%) with a very high next-hop router diversity (>=30).

Intuitively, the number of visible next-hop routers increases as the number of unique neighboring AS increases. We further find that the increase in the amount of visible next-hop routers mainly depends on the type of the neighboring AS through which ISP_A reaches a given destination.

Figure 5 shows the number of peering routers for different types of neighboring ASes. The types of neighboring AS are based on the classification found in [25], Figure 5 indicates that in general larger neighboring ISPs tend to have a higher number of routers peering with ISP_A. This tendency is reflected in next-hop diversity and is the main reason behind the next-hop diversity differences between different prefixes. For example, if two prefixes are reached via one Tier1 and a small ISP neighbor respectively, then based on Figure 5, the

former prefix can potentially have its next-hop router diversity ranging from 6 to 12 while the diversity of the latter prefix can only range from 1 to 8. Last, we observe that a few stub neighbor ASes (ex., UltraDNS, Amazon, etc) that have exceptionally high number of peering routers. These ASes connect to ISP_A with multiple routers, but they are classified as "stub" given that they do not provide transit service.

Note that our observation of next-hop diversity shows important evidence that, without modification to BGP, there exist opportunities in ISP_A's current network for fast failure recovery (multiple next-hop routers to reach a given prefix), traffic engineering (multiple PoPs to reach a prefix), and load balancing (multiple next-hop routers in a same PoP).

B. Case Studies

In this section, we take a closer look at representative cases of prefixes with the low, moderate, and high next-hop diversity to understand the main factors that determine the amount of next-hop diversity for a given prefix.

1) Low Diversity: In this case, prefixes have low diversity. For example, prefix 201.133.104.0/24 announced by AS8151 has one next-hop. AS8151 directly connects to ISP_A, and the number of observed AS level topology is shown in Figure 6(a). Based on this static observation, there can be two reasons why this prefix has the lowest next-hop diversity: 1) there is only one path to reach the prefix through only one next-hop router, and/or 2) BGP's design choice to select and propagate only the selected path to the neighbors prevents ISP_A from being able to see other alternative paths. To understand which one is responsible for the lowest diversity of these prefixes, we further investigate the update messages and find that the main reason is the latter.

This is, when the best (and the only visible) path fails, we could observe that other alternative paths got exposed during iBGP convergence process; these paths were hidden from the BGP routers because they were not selected as the best path and were not propagated further to the neighboring routers.

Using AS level Internet topology available from [25], we verified that for all prefixes with next-hop router diversity equal to one, they do have multiple alternative next-hop routers, which were hidden from the BGP routers when the best path was stable. This observation suggests that when BGP is modified such that the less preferred paths are not hidden, these prefixes with the lowest path diversity will benefit the most.

Observation Summary: Although alternative paths do exist, BGP's design choice to select and propagate only a single best path hides the alternative paths and prevented the prefixes in this class to have higher diversity.

2) Moderate Diversity: Figure 4 shows that there are more than 47% prefixes have their next-hop diversity between 6 and 11, and we refer to prefixes with next-hop diversity from 6 to 11 as prefixes with moderate diversity.

There are two representative cases of prefixes with a moderate next-hop diversity. Prefix 190.103.225.0/24 announced by AS27983 is the first case. This prefix can be reached from ISP_A through AS6762, a large ISP. The number of next-hop routers between ISP_A and AS6762 were 7. Another representative case of a prefix with moderate next-hop diversity was prefix 204.113.217.0/24 announced by AS210. The AS path and next-hop diversity are 2 and 12 respectively.

In both examples, the prefixes were reached through at least one neighbor AS which is a large ISP and has at least 6 BGP peering sessions with ISP_A

Observation Summary: Between ISP_A and other large ISPs, there are in general 6 to 11 BGP peering points across the globe. Therefore, the prefixes reached via the neighboring large ISPs have their next-hop diversity from 6 to 11 at least.

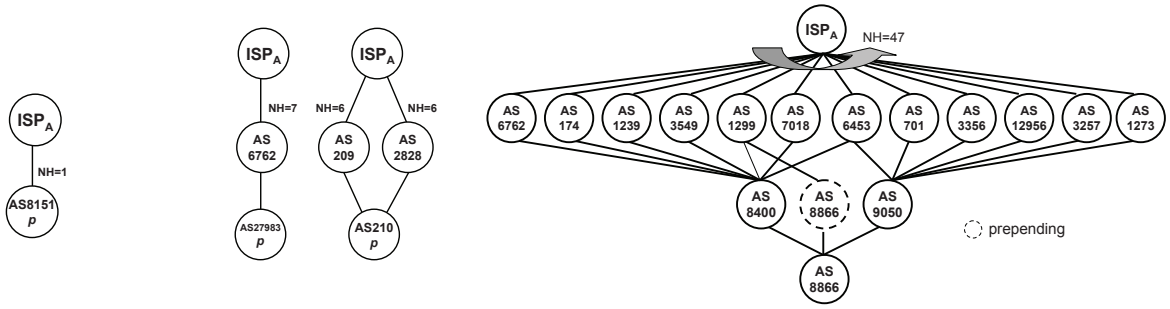
3) High Diversity: In this section, we present two prefixes with the high degree of next-hop diversity. The first prefix we present is 83.228.80.0/23 announced by AS8866. AS8866 is a regional ISP, which further multi-homes with different providers who are highly connected to many Tier1 ISPs. The two providers are: AS8400 and AS9050. These two providers are also telecommunication service providers themselves, and they are customers of 6 or 7 different large ISPs or Tier1s ISPs as shown in Figure 6(c). By becoming a customer of these two highly connected providers, prefix 83.228.80.0/23 in AS8866 inherently becomes visible through highly diverse paths from the perspective of ISP_A. Note that all AS paths from the origin AS to ISP_A happened to be equal in length.

Because of each backbone router in ISP_A prefers the eBGP learned path over the other iBGP learned path, there was no "hidden path phenomenon" in this case. If the origin AS had any path with shorter AS path length, all alternative paths would have been withdrawn and hidden from other backbone routers.

Our second example of a prefix with high next-hop diversity is prefix 64.94.107.0/24 announced by AS27281 (not shown in the figure). From the view of ISP_A, this prefix could be reached using 28 distinct AS neighbors and 40 next-hop routers from ISP_A. We checked that AS27281 provides service for web measurements and user profiling, and surprisingly, it has only one provider, Internap [5]. Internap has multiple sibling AS numbers, and use different AS numbers to connect to other providers in multiple PoPs. As a consequence, Internap's customers could naturally inherit high number of different AS paths.

This example reveals an interesting fact that one can achieve higher degree of path diversity by using different AS numbers and announce via diverse paths. However, we note here that the number of prefixes who *intentionally* maintain high diversity with ISP_A are very small.

Another interesting common characteristic of prefixes with high degrees of next-hop diversity is that their origin ASes do not directly connect to ISP_A. If the origin AS later establishes a BGP connection such that the newly established BGP path is more preferable based on the BGP best path selection rule (*e.g.* a route through a customer and/or with shorter AS path



(a) Case 1: Low Diversity (b) Case 2: Moderate Diversity (c) Case 3: High Diversity

Fig. 6. Representative Cases of Prefixes with Low, Moderate, and High Next-hop Diversity

length), then this new AS path will be selected as the new best path. As a consequence, all other alternative AS paths that were previously visible with their length greater than one will be withdrawn due to the path poisoning phenomenon explained in Section II-D.

For example in Figure 6(c), assume that AS8866 peers directly with ISP_A and starts to announced prefix 83.228.80.0/23. This newly announced AS path has its length equal to one and will be preferred over other previously visible paths with their AS path length equal to two. As a result, the previously visible path with longer AS paths will be withdrawn and hidden, and this is an unavoidable consequence of the current BGP design to select and propagate only a single best path.

Observation Summary: The main reason that the prefixes maintain the high degree of next-hop diversity is that the prefixes are announced to ISP_A using multiple next-hop routers from multiple large neighboring ASes. In most cases, the high degree of nexthop diversity is due to topological connectivity: origin ASes do not directly peer with ISP_A, and they multi-home with other large ISPs, or their providers multihome with large ISPs. These providers then happen to peer with ISP_A directly or indirectly with equal-length AS paths. Therefore, all paths and the corresponding next-hop routers are visible from ISP_A, leading to the high degree of next-hop diversity.

C. Lack of Geographical Presence and High Diversity

From the above case studies, we found that the prefixes with the high degree of next-hop diversity share one interesting property: their origin ASes do not connect directly with ISP_A during our measurement period. Rather, the origin AS multi-homes with a few large providers, and in turn these providers connect to ISP_A using multiple next-hop routers except ISP_A.

From this observation, we hypothesized that the lack of geographical presence of ISP_A can be a factor that determined the set of high next-hop diversity prefixes. In the regions that

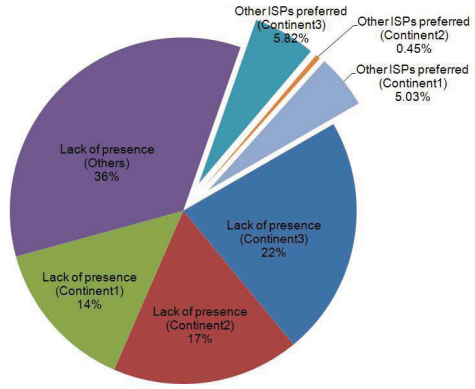


Fig. 7. Geographical Presence of ISP_A

ISP_A does not provide connectivity, the origin ASes have to connect to the local ISPs when they wish to connect to the Internet. If these local ISPs happen to multi-home with many large ISPs except ISP_A, then there will be many paths with equal AS_PATH length between the origin AS and ISP_A. As a result, this leads to a very high number of next-hop routers visible from ISP_A to reach the prefix.

To verify our hypothesis, we checked the prefix origination point of prefixes with very high next-hop diversity against the PoPs covered by ISP_A. To find the location of prefix origination point, we used MaxMind GeoLite package [2] to map each prefix into a city. Then for these cities, we checked whether any PoP of ISP_A is present. Figure 7 verifies our hypothesis. In 89% of prefixes with very high next-hop diversity, ISP_A did not have a presence. In case of the rest 11% of prefixes, the origin ASes do not directly peer with ISP_A but are connected to other ISPs. From this observation, we conjecture that the set of prefixes with very high next-hop diversity will differ from one ISP to another ISP, mainly due to the geographical coverage difference between the ISPs.

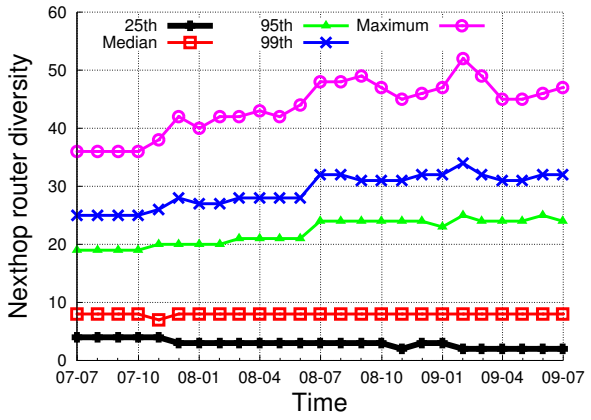


Fig. 8. Next-hop Diversity Change in Time

V. Trends of Next-hop Diversity in Time

In the previous section, we quantified and analyzed BGP next-hop diversity from ISP_A at a given time instance. In this section, we seek to find out if there is a general trend of next-hop diversity changes over time.

Due to a large amount of iBGP routing data and the processing loads, we calculated next-hop diversity for the routing table snapshot (RIB) on the first day of each month from July 2007 to July 2009. In addition, to better capture the next-hop diversity change for a given prefix, we only consider the prefixes that continuously exist over the entire two-year measurement period. The total number of such prefixes across all the RIBs is 220,432.

Figure 8 depicts next-hop router diversity changes at 25, 50, 95, 99 percentile, and maximum in next-hop router diversity distribution curves at different times. For example, on July 2007, the median, 99%, and maximum next-hop diversity were 8, 25 and 36, while on July 2009, next-hop router diversity were 8, 31 and 48, respectively.

Figure 8 shows that over the last two years, maximum, 99 percentile, and 95 percentile next-hop router diversity gradually increased in time, whereas the median value stayed the same, and 25 percentile value decreased slightly. After further investigation, we found that the increasing trend in maximum, 99 percentile, and 95 percentile next-hop router diversity is mainly due to the increased number of peering routers between ISP_A and its neighbors. Since July 2007, the number of backbone routers in ISP_A gradually increased and had up to 19 additional routers by the end of July 2009.

VI. Related work

Prior works on path diversity fall into two classes: 1) quantifying existing path diversity and 2) increasing path diversity.

In the first class, [20, 13, 3, 23, 6, 22] attempt to quantify and understand the path diversity in the view of different domains and network levels. Teixeira *et al.* [20] measure the IP level path diversity inside a Tier1 ISP (Sprint)'s backbone network. By using IGP routing data to reconstruct the underlying backbone topology between all PoP (Point of Presence) pairs, their results show that Sprint has significant *IP level* path diversity among their PoPs. In contrast, we measure the *BGP level* exiting point diversity. The difference is that even though the underlying IGP topology may provide different IP level paths, such diversity could be mapped to the same exiting (*i.e.* , nexthop) routers at the BGP level.

There exist some prior measurement studies that focus on quantifying the path diversity at the BGP level. In [13, 3, 23, 6], the authors measure path diversity at the AS granularity with similar methodologies as described in Section III. However, their common goal is to understand the impact of path diversity on data forwarding performance for a given multi-homing AS. As the result, their studies mostly focus on multi-homing stub ASes, while our work measures path diversity from the perspective of a Tier1 ISP.

As one of the most closely related to our work, Uhlig *et al.* [22] quantifies path diversity in a Tier1 ISP which configured its network with Route Reflection. They simulate 1,000 prefixes with the highest amount of traffic volume. Because their main goal is to gain an insight on the impact of Route Reflection on path diversity, they used normalized metrics that represent path diversity. Our work, on the other hand, measures next-hop diversity in a full-mesh network, and quantifies path diversity for all prefixes in the global routing table using the actual number of routers, which yield a more tangible and comprehensive understanding of both general and corner cases.

The second class of prior works involve efforts to increase path diversity. Recently, the operator community started to demand higher path diversity to satisfy requirements for the newly emerging applications [17, 9, 10]. This led to several on-going efforts to increase path diversity by modifying the behavior of BGP. Raszuk *et al.* [17] proposed to modify BGP such that multiple paths can be distributed instead of the single best path. By doing so, other BGP peers can learn multiple paths to reach a given destination. Walton *et al.* and Schrieck *et al.* [9, 10] propose and analyze a new mechanism to increase path diversity by distributing multiple paths for a given destination. In our work, we quantify the amount of existing path diversity in a Tier1 ISP, and can serve as an objective evidence to decide whether such mechanisms to increase path diversity are necessary.

Lastly, a few prior works focus on analyzing and determining the optimized best path (with regard to data forwarding performance, traffic engineering, quality of service, and etc) assuming that multiple paths are available. Buob *et al.* [8] proposes an optimization scheme to select the best path amongst multiple paths to a given destination. Our work provides details on what and how many prefixes can really benefit from such optimization via an operational Tier1 ISP's perspective.

VII. Discussion and Future works

In this work, we quantify next-hop diversity in the view of a Tier1 ISP, which represent a perspective from large ISPs with

full-mesh iBGP configuration. We identify three factors that could have significant impacts on a prefix's next-hop diversity.

First, as shown earlier in this work, the number of next-hop routers and the type of neighboring ASes play an important role on the observed next-hop diversity, and therefore, next-hop diversity in ISPs of different sizes may be different than that of ISP_A studied in this work.

Second, it is known that full-mesh configuration preserves the optimal path diversity, and other iBGP configurations such as Route Reflector may yield different results. Thus, understand the path diversity in Ases with route reflectors is subject to our ongoing work.

The third factor that can potentially affect the observed diversity is the relationship between an AS and its neighbors, and how prefixes are announced by different types of neighbors. Because BGP prefers routes from certain types of neighbors (*e.g.* prefer customer routes over peer routes), the less preferred routes can be hidden due to "hidden path phenomenon" described earlier in this paper in Section II-D. Because the number of neighbor ASes and their relationships differ from one network to another, the impact may differ as well. Measuring the number of routes announced over BGP sessions of different relationships and their impact on the path diversity can be interesting.

Last, another issue concerns the impact of next-hop diversity on the dynamic routing convergence. One limitation in this work is that we focused on understanding the static BGP diversity. Although we show that the majority of prefixes have multiple exiting next-hop routers, which shall be qualitatively more robust to the internal errors, it remains an open question that how fast BGP can converge after failures: does it correlate with the number of next-hop routers, or one backup next-hop is suffice? Along with our study of static BGP path diversity, we observed that the number of BGP update message exchanges and the convergence delay can be different based on next-hop diversity for a given prefix. More measurement studies are necessary to quantify and understand the impact of BGP path diversity on convergence time.

VIII. Summary

BGP has gone through many changes as it operates as the de-facto routing protocol in the Internet. Its original design required a BGP router to select and propagate only a single best path to its neighbors. Advancement in both hardware and software has enabled a router to scale better and recently in IETF, this design choice is being reconsidered to improve the robustness, flexibility in traffic engineering, and convergence. However, there has been little understanding on path diversity in the existing system, and the necessity and effectiveness of different proposals are not clear.

Using iBGP routing data collected from more than one hundred backbone production routers, our results show that the majority of prefixes could be reached through multiple next-hop routers. There exist 88% of prefixes which could be reached via at least 2 next-hop routers. Our analysis shows that local routing preference and the number of peering points are the two dominating factors in general. Furthermore, we find that a very small number of prefixes maintain a high degree of diversity, and in most cases, they happen specifically by the lack of geographical presence of ISP_A in the regions where origin ASes are located.

From the data of two recent years, we observe an interesting fact that the overall next-hop diversity have not changed much, while individual prefix does shift its diversity to some extent. As the individual prefix's diversity is determined by a complex interaction between the topological and geographical location of the origin AS, the inter-domain routing path from the origin to ISP_A, the number of next-hop routers, and the BGP routing decisions, the path diversity changes in time with a seemingly unpredictable manner. However, we observed that the maximal next-hop diversity slowly increases in time, mainly due to the increased number of backbone routers inside ISP_A.

References

[1] BGP Parser. http://irl.cs.ucla.edu/bgpparser/.

[2] MaxMind - GeoIP. http://www.maxmind.com/app/ip-location.

[3] Measuring Provider Path Diversity from Traceroute Data: work in progress. http://www.caida.org/outreach/isma/0112/talks/krishna/index.pdf.

[4] MRT Routing Information Export Format. http://www.ietf.org/internet-drafts/draft-ietf-grow-mrt-07.txt.

[5] The Internap Network-Based Route Optimization Solution. http://www.internap.com.

[6] P. D. Arjona Villicana, C. C. Constantinou, and A. S. Stepanenko. The Internet's Unexploited Path Diversity, 2009.

[7] T. Bates, E. Chen, and R. Chandra. RFC 3345: BGP Route Reflection: An Alternative to Full Mesh Internal BGP.

[8] M.-O. Buob, M. Meulle, and S. Uhlig. Checking for Optimal Egress Points in iBGP Routing. *Design and Reliable Communication Networks*, October 2007.

[9] A. R. D. Walton and E. Chen. Advertisement of Multiple Paths in BGP, March 2010.

[10] V. V. den Schrieck and P. Francois. Analysis of Paths Selection Modes for Add-paths, July 2009.

[11] A. Flavel and M. Roughan. Stable and Flexible BGP. In *ACM Sigcomm*, October 2009.

[12] S. Halabi and D. McPherson. Internet Routing Architectures, 2nd ed., August 2000.

[13] J. Han, D. Watson, and F. Jahanian. An Experimental Study of Internet Path Diversity. In *IEEE Transactions on Dependable and Secure Computing*, October 2006.

[14] N. Kushman, S. Kandula, and D. Katabi. Can You Hear Me Now?!: It Must Be BGP. *SIGCOMM Comput. Commun. Rev. (CCR)*, 37(2):75–84, 2007.

[15] C. Labovitz, A. Ahuja, A. Abose, and F. Jahanian. Delayed Internet Routing Convergence. *IEEE/ACM Transactions on Networking*, 9(3):293 – 306, June 2001.

[16] P. Marques, R. Fernando, E. Chen, and P. Mohapatra. Advertisement of the Best External Route in BGP, Febrary 2010.

[17] R. Raszuk, K. Patel, I. Kouvelas, R. Fernando, and D. McPherson. Distribution of Diverse BGP Paths, March 2010.

[18] Y. Rekhter, T. Li, and S. Hares. RFC 4271: A Border Gateway Protocol 4 (BGP4), January 2006.

[19] Y. Schwartz, Y. Shavitt, and U. Weinsberg. A Measurement Study of The Origins of End-to-End Delay Variations. In *ACM SIGCOMM Passive and Active Measurement Conference (PAM)*, 2010.

[20] R. Teixeira, K. Marzullo, S. Savage, and G. M. Voelker. In search of path diversity in ISP networks. *ACM Sigcomm conference on Internet measurement*, 2003.

[21] P. Traina, D. McPherson, and J. Scudder. RFC 5065: Autonomous System Confederations for BGP, August 2007.

[22] S. Uhlig and S. Tandel. Quantifying the BGP Routes Diversity Inside a Tier-1 Network. *Networking*, 3976, April 2006.

[23] V. Vasudevan, D. G. Andersen, and H. Zhang. Understanding the AS-level Path Disjointness Provided by Multi-homing, 2007.

[24] F. Wang, Z. M. Mao, L. G. Jia Wang, , and R. Bush. A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance. In *Proceedings of ACM SIGCOMM 2006*, 2006.

[25] B. Zhang, R. Liu, D. Massey, and L. Zhang. Internet Topology Project. http://irl.cs.ucla.edu/topology/.