The Hong Kong Polytechnic University

Department of Computing

Measuring Routing Dynamics Induced by the AS Path Prepending Method

Samantha Sau-Man Lo

A thesis submitted in partial fulfillment of the requirements for the Degree of Master of Philosophy

October 2007

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Sau-Man Lo (Name of student)

Abstract

Thousands of autonomous systems (ASes) connect to each other to form the Internet today. They exchange reachability information via the only inter-domain routing protocol, Border Gateway Protocol (BGP). BGP provides attributes for individual ASes to express their routing preferences which are often a result of the routing policies. Moreover, an increasing number of ASes connect to more than one AS in order to facilitate multi-homing. Thus, inbound traffic engineering has become a crucial task for network operators. Among a handful of inter-domain inbound traffic engineering methods, AS path prepending is a widely practiced method which provides network resilience and does not increase routing table size. Unfortunately, network operators often perform prepending on a trial-and-error basis, which can lead to suboptimal results and a large amount of network churn.

This dissertation studies the effects of the AS path prepending method based on Internet measurement. In particular, we have developed an active measurement methodology to study the prepending method. In this method, we actively inject prepended routes into the Internet routing system. We have implemented this method in two networks. One is on the RIPE Network Coordination Centre (NCC) Routing Information Service (RIS) which is one of five Regional Internet Registries (RIRs). The other is a Hong Kong local campus network. We have observed the resulting changes from almost 200 publicly-accessible sources of BGP information. Our results show that the measurement methodology is scalable and effective to study the effects of prepending announced by a stub AS which uses prepending to control its inbound traffic. Our analysis also shows that a small number of ASes is often responsible for a large amount of route changes induced by prepending. Furthermore, our methods are able to reveal hidden prepending policies and tie-breaking decisions made by ASes. These new observations and insights are useful for further predicting the effectiveness of the prepending method.

Publications Arising from the Dissertation

1. Conference Publications and Presentations

- LO, S. and CHANG, R. K. C. (with COLITTI, L.) An Active Approach to Measuring Routing Dynamics Induced by Autonomous Systems. Proceedings of ACM FCRC Workshop of Experimental Computer Science (ExpCS), San Diego, pp. 21-30, June 2007.
- LO, S. and CHANG, R. K. C. Measuring the Effects of Route Prepending for Stub Autonomous Systems. Proceedings of ICC Workshop on Traffic Engineering in Next Generation IP Networks, Glasgow, June 2007.
- LO, S. and CHANG, R. K. C. Active Measurement of the AS Path Prepending Method. The North America Network Operators' Group (NANOG 37) (Talk), San Diego, June 2006.
- LO, S. and CHANG, R. K. C. Active Measurement of the AS Path Prepending Method. IEEE International Conference Network Protocols (ICNP) (poster paper), Boston, November 2005.

2. Other Publications

- LO, S. and CHANG, R. K. C. (with COLITTI, L.) RIPE RRC07 beacon AS paths with AS prepending. Internet Measurement Data Catalog (DatCat), San Diego, March 2007.

  `http://imdc.datcat.org/collection/1-01J5-7=RIPE+RRC07+beacon+`

  `AS+paths+with+AS+prepending;jsessionid=54865662414C24D4D9398AD1F1DAC0C0`

- LO, S. and CHANG, R. K. C. (with COLITTI, L.) RIPE RRC10 beacon AS paths with AS prepending. Internet Measurement Data Catalog (DatCat), San Diego, March 2007. http://imdc.datcat.org/collection/1-01J6-V=RIPE+RRC10+beacon+AS+paths+with+AS+prepending;jsessionid=54865662414C24D4D9398AD1F1DAC0C0

- LO, S. and CHANG, R. K. C. (with COLITTI, L.) RIPE RRC14 beacon AS paths with AS prepending. Internet Measurement Data Catalog (DatCat), San Diego, March 2007. http://imdc.datcat.org/collection/1-01J7-G=RIPE+RRC14+beacon+AS+paths+with+AS+prepending;jsessionid=54865662414C24D4D9398AD1F1DAC0C0

Acknowledgments

I would like express my deepest gratitude to my research supervisor Prof. Rocky K. C. Chang for his invaluable support, advice, insight, and guidance throughout this interesting and challenging research project.

I am very grateful for Prof. Nick Feamster and his suggestions and insightful ideas about this research work while I was a visiting student at the Georgia Institute of Technology.

I also thank Mr. Michael Lo for providing measurement facilities and giving suggestions to this work.

I thank Dr. Lorenzo Colitti from Google (formerly from RIPE) for helping me with measurement setup and discussions. This dissertation has been significantly enriched through their collaborations and help.

In addition, I would like to thank my great research team members, Dr. Daniel Luo, Mr. Edmond Chan, Ms. Grace Xie, Ms. Kathy Tang, and Mr. Sam Lam for sharing their suggestions, experiences, and knowledge with me.

I want to also thank The Hong Kong Polytechnic University for their generous support for my six-month visit at the College of Computing, Georgia Tech.

Last but not the least, I owe everything to my family - my parents and relatives for their endless love, support, and encouragement. This thesis is

dedicated to all of them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Internet is a network of hundreds of millions of nodes that exchange information with each other. Because of the dynamic nature and sheer size of the network, routing of data within it takes place on two hierarchical levels. At the highest level, the Internet is partitioned into tens of thousands of administrative domains known as autonomous systems (ASes). Each AS is identified by an integer number and corresponds to an organization which can be an Internet Service Provider (ISP), government or campus network. ASes cooperate with their neighbors to ensure the global reach of all destinations on the network (*inter-domain routing*). At the lowest level, each AS has a complete view of its own network and is responsible for the routing of Internet traffic within its boundaries (*intra-domain routing*).

Inter-domain routing is performed by a single inter-domain routing protocol, Border Gateway Protocol (BGP) [43]. BGP is used for ASes to exchange reachability information and apply routing policies to their networks to facilitate Internet traffic engineering purposes. Inter-domain routing is considered much more difficult than intra-domain routing since inter-domain routing requires tens of thousands of independently-operated ASes to cooperate to route traffic. Increasing the number of ASes with unknown topology and routing policies complicates inter-domain routing.

There are more than 25,000 ASes in the Internet [13] and the numbers of ASes and multi-homed ASes are increasing. In 2004, more than 50% of ASes were stub multi-homed ASes [10]. Stub ASes subscribe the Internet service from transit ASes, and they do not carry traffic for other ASes. Transit ASes, on the other hand, carry traffic on behalf of other ASes. Both multi-homed and transit ASes have more than one link to the Internet. Furthermore, such an AS could establish multiple links with an upstream provider. These ASes can also have more than one upstream AS. Each AS operates independently and does not reveal its topology to other ASes.

Inter-domain routing is challenging to network operators because every AS can apply its own routing policies without noticing other ASes. Multi-homed ASes apply routing policies to control their traffic. They can specify the preference of link to be used to send traffic without notifying the destination AS. They can use one of the links as a backup link and another one as a primary link [1]. Although the backup link may not have a high performance, when the primary link breaks down, the backup link can replace the primary link. Multi-homed ASes can also distribute the traffic between two or more link to optimize the traffic and the cost of the links. These configurations are

[1] Here a link can be a physical link between the multi-homed AS and its upstream provider, or just a connection between the multi-homed AS and its upstream provider.

unknown to other ASes. The unknown topology and routing policies lead to difficulties of inter-domain routing. As a result, traffic engineering on inter-domain routing is complicated.

Optimizing traffic and utilizing resources, Internet traffic engineering [5] enhances the performance of an operational network. Network operators control the amount of traffic on a link according to the performance requirements. Performance of a link is affected by delay, delay jitter, packet lost rate, throughput, and packet re-ordering. At the same time, Internet traffic engineering requires network operators utilize network resources economically and reliably. Network resources of a link include link bandwidth, buffer space, and computational resources. Thus, Internet traffic engineering is important to operational networks.

Internet traffic engineering can be divided into intra-domain and inter-domain traffic engineering. Intra-domain traffic engineering usually is done on layer two or layer three (IP) within an AS. IP-based techniques are performed on an IP network by adjusting the attributes of intra-domain routing protocol, or Interior Gateway Protocol (IGP). Several common IGPs include OSPF [36] and IS-IS [26]. Another technique performing intra-domain traffic engineering, MPLS (Multiprotocol Label Switching) [20] technique, uses label-switching or frame-relay networks to determine the outgoing interface of a packet. In label-switching, a switch puts a label on a packet and thereby allows the router to judge which interface should be the outgoing interface. This technique enables the network operator to adjust the traffic load of that interface. As a result, intra-domain traffic engineering is under the control of network operators in the same AS through all these techniques.

In contrast to intra-domain traffic engineering, inter-domain traffic engineering involves a large number of ASes to control traffic. Each AS has its own routing policy, and manages its inbound and outbound traffic. Inter-domain traffic engineering can be categorized into inbound and outbound traffic engineering. Inbound traffic engineering controls how traffic is sent into a network. Outbound traffic engineering controls how traffic sent out from a network. Inbound traffic engineering is comparatively more complicated than outbound traffic engineering. In outbound traffic engineering, the traffic initiator itself decides the way the traffic is sent to the destination network. On the other hand, inbound traffic engineering relies on several methods to change the routing decision of the traffic initiator, and many intermediate ASes' decisions are difficult to predict.

Network operators use several methods to perform inbound traffic engineering. The methods can be divided into BGP level and application level. On the BGP level, selective announcement, specific prefix advertisement, Multi-Exit Discriminator (MED), BGP communities, and AS path prepending are mainly used to control inbound traffic of multi-homed ASes. Selective announcement and specific prefix advertisement allow operators to distribute the incoming traffic of one prefix across different links. However, they increase the routing table size of routers by introducing more detail prefixes to the routing tables. Moreover, in some cases, they do not provide network resilience. If one of the connections is not available, the network operator needs to announce the prefix manually.

Another method to perform inbound traffic engineering on the BGP level is Multi-Exit Discriminator. Unlike specific prefix advertisement, MED can be only used when more than one link exists between two ASes to control the distribution of traffic among these links. Therefore, the MED attribute cannot be used to control the traffic from more than one provider or upstream AS.

BGP communities can also be used to perform inbound traffic engineering on the BGP level. Network operators of an AS can use these community values to control the routing policies in the upstream providers of the AS. When the routers of the upstream providers receive the community values attached on a BGP advertisement, the routers perform some actions according to the assignments in the route-map. These actions include assigning local preferences, prepending, or changing other attributes in that BGP advertisement in order to change the flow of traffic. Therefore, the underlying method is still prepending. However, the BGP community allows you to perform prepending in the upstream network. Still, special coordination between ASes is required for using communities to control traffic.

On the application level, dynamic NAT-based traffic engineering is also used to tune the inbound traffic with BGP [28]. A gateway implements Network Address Translation (NAT) and replaces the source IP address of outgoing packets with an address of a NAT box according to internal load-balancing policy on that NAT box. However, the scalability of this method has been questioned when NAT boxes are overloaded or implemented on a large ISP.

Unlike other inter-domain inbound traffic engineering methods, AS path prepending (prepending) provides network resilience and does not increase routing table size. The prepending method inflats the AS path by adding multiple of its AS number into the AS path attribute. Operators usually use this method to discourage the upstream providers to choose the link. In earlier studies, Feamster et al. [21] and Broido et al. [8] reported their findings on prepending based on the AT&T backbone data and RouteViews data, respectively. In a recent passive measurement study, Wang et al. [51] provided a more up-to-date report using various statistics compiled from the RouteViews data and proposed a model to study fundamental issues of decentralized traffic engineering in the Internet. These studies show that prepending is very common for network operators to control inbound traffic.

Unfortunately, the effects of prepending can be difficult to predict. One problem in anticipating the effects of prepending is that the results depend on the routing policies of different ASes, policies we can't know. Routing policies include preferring a route with higher local preference, a shorter AS path, or other metrics. An AS chooses its own local preferences for which link the AS would like to use to send traffic to the destination network. If an AS assigns the same local preference value to all the links, the AS considers the AS path length to each individual destination network. The destination can perform prepending to inflate the AS path length, but it cannot directly control other ASes' routing polices. Because the destination network does not know the preferences of other ASes and whether they consider the AS path length, it cannot predict the outcome of prepending.

A second problem in predicting the effect of prepending is unknown network topology. Not all ASes are multi-homed or have both backup and primary connections available to them. Upstream ASes of these ASes decide the routing paths for them, and these ASes cannot control the way of sending the traffic. Furthermore, the amount of traffic sent from different ASes does not depend on the number of IP addresses they have. Some of the IP addresses can send more traffic which depends on the service they provide. Operators do not know how much prepending length we should add in order to shift a certain amount of traffic. Because of that, network operators often perform prepending on a trial-and-error basis, which can lead to suboptimal results and delayed routing convergence.

Internet routing convergence is another issue for inter-domain routing. The effect of prepending takes time to converge [53] because of the route flap damping mechanism in BGP, which is used to stabilize the routing path under frequent BGP announcement. The route convergence of prepending has not been studied. Furthermore, routing events include routing policies, iBGP configuration and MRAI timer values are proven to degrade end-to-end Internet path performance [50]. The effects of prepending to end-to-end Internet path performance are still unknown. Without understanding of the effect of prepending, performance of the Internet is further degraded by trial-and-error prepending operations.

The previous work [51] attempted to study the behavior of prepending by passive measurement. Passive measurement uses publicly available BGP routing data archives to study prepending. In [51], Wang et al. identified the prepended route from the archives. However, because the purpose of prepending is to discourage the use of a given route, routers are not expected to choose the prepended route as a best route. Then the routing data will not contain these non-best routes. Thus, we cannot measure the effect of prepending by passive measurement. However, we can measure prepending by announcing prepended routes and observing their effects, i.e. active measurement.

In terms of active approaches, some previous works only focused on measuring the traffic shift induced by prepending. Chang and Lo [12] proposed to use `AutoPrepend`, an automated procedure, to determine the best prepending length without using trial-and-error. A key component of `AutoPrepend` was a beacon prefix to predict the magnitude of the incoming traffic variation due to prepending. They changed the prepending length of the beacon prefix and observed the link shift by generating traffic to different top traffic senders. Quoitin et al. [41] conducted similar active measurement experiments to study prepending. However, Quoitin et al. only studied the incoming traffic volume shift without understanding the factors of the traffic shift on the routing level. Moreover, their methods generated traffic on the production network. In order to understand the effects of prepending to a certain network or topology, we must study prepending on the routing level.

1.1 Motivation

Routing dynamics induced by ASes are difficult to predict without the knowledge of the routing policies of ASes and topology of the Internet. The routing policies of each individual AS decides which path the traffic would be sent through. Multi-homing affects the number of possible routes that are available to ASes. Unfortunately, it is impossible to get the routing policies of all the ASes in the Internet which are considered to be confidential. Furthermore, the current way of studying the topology cannot reveal all the possible paths in the Internet.

As a result, when network operators perform traffic engineering or normal operations, they cannot directly predict the results of that operation. They observe the traffic shift and then decide the suitable method to use without any understanding of the effects. Prepending is one of the common operations whose effects are difficult to understand. The effects of prepending are not deterministic because they depend on a complex interaction of the upstream ASes' routing policies. There is a lack of understanding of convergence after they announce prepending to the Internet, and furthermore routing dynamics may degrade the end-to-end path performance during the period.

1.2 Problem Statements

In this dissertation, several key issues on routing dynamics induced by ASes are addressed and studied:

1. Why is AS path prepending effective to control inbound traffic?

   Although AS path prepending is effective at controlling the inbound traffic, we still don't know the reason for this effectiveness. Current studies focus on how prepending shifts traffic instead of the reasons that prepending is effective. Without understanding how prepending controls the inbound traffic, we can only perform prepending on a trial-and-error basis.

2. How can we measure the routing dynamics induced by AS path prepending?

Currently available methodologies, passive and active measurements, cannot truly allow us to observe the routing dynamics that AS path prepending induces. Without information about these routing dynamics, we cannot understand why prepending changes traffic volume.

3. What are the causes AS path prepending to affect inbound traffic volume?

   It is difficult to predict the routing dynamics induced by prepending. Once a BGP router receives a BGP announcement, it decides to use the routing information on the newly received BGP update based on the routing policy. Then the BGP router propagates the new information to its neighbor routers and peer ASes based on the routing policy. However, routing policies are not publicly available.

   Because we can't know the topology and routing policy of ASes, routing dynamics are not predictable and it is difficult to study the behavior of prepending. Many operators don't know these factors affecting the behavior of prepending, so they have to use it on a trial-and-error basis. This guesswork can increase the time of convergence as well as the router's workload. Understanding of these factors would be highly helpful for understanding the effects of prepending and improve the process of daily network operations.

4. How can we help network operators perform AS path prepending?

   Because network operators perform prepending on a trial-and-error basis, there is a need for tools to help them perform prepending.

1.3 Contributions

In this dissertation, we address several key research issues concerning the routing dynamics induced by ASes:

1. This is the first intensive study of AS path prepending. We have proposed an active measurement methodology for studying the problems described above. We have provided a framework for studying the routing dynamics induced by inbound traffic engineering. No real traffic is involved, and the measurement is simple to deploy. Every AS can use this method to study the effect of prepending on their network. In this dissertation, we have used the beacon prefix to study the route convergence issues for route announcement with different prepending lengths.

2. We have implemented this methodology in two networks, a Hong Kong university's local network and the RIPE Network Coordinate Centre[44] in Europe. We study the responses of these networks to prepending.

3. We have determined and classified the upstream ASes according to their responses to prepending. We have analyzed the route changes induced by prepending and classify ASes according to their responses to prepending. This classification is useful to identify upstream ASes and study their behaviors under prepending.

4. In our result, we have found that a small number of ASes are responsible for a large amount of route changes. These ASes have a large number of downstream ASes, and changes in the route traffic of the downstream ASes follow the changes in the traffic of the smaller group of ASes. With

the identification of the responsible ASes, we can understand the reason for sudden shifts of traffic when network operators perform prepending.

5. The measurement data, the first of its kind, has been archived and indexed on DatCat.org [19].

1.4 Organization of This Dissertation

- Chapter 2 gives background about inbound traffic engineering and related works about studying AS path prepending include passive and active measurements, and methodologies to determine optimal prepending length.
- Chapter 3 presents the active measurement methodology.
- In Chapter 4, we further describe the implementations on the two networks, and the difficulties and constraints of the implementations.
- Chapter 5 presents the results and analysis.
- Chapter 6 concludes the dissertation and gives promising future works for AS path prepending.

Chapter 2

Literature Review

In this chapter, we present background of routing and important related works in this area. We first introduce inter-domain routing and inbound traffic engineering. We will introduce passive and active measurement methodologies, and give you idea on why only passive measurement is not enough for studying routing dynamics. We will further talk about some available tools for the measurement and some related works.

2.1 Inter-domain Routing

The Internet is composed of IP (Internet Protocol) networks. Each IP network is a contiguous block of IP addresses and we call each IP network a *prefix*. A collection of prefixes and routers that are under the same control of a single domain and presents a common routing policy to the Internet is an Autonomous System (AS). For example, The Hong Kong Polytechnic University is an AS. An Internet Service Provider (ISP) is also an AS. An AS has its own AS number to identify itself. The Regional Internet Registries are responsible for managing the AS numbers and IP addresses on the Internet [37]. They are the African Network Information Center (AfriNIC) [1], the American Registry for Internet Number (ARIN) [2], the Asia Pacific Network Information Center (APNIC) [3], the Latin American and Caribbean IP Address Regional Registry (LACNIC) [31], and the Reseaux IP Europeans (RIPE) [44].

ASes are classified into transit, stub, single-homed, and multi-homed ASes. The classification is based on their functionalities. If an AS provides connectivity service to another AS, for example, an ISP, and carries the traffic for its customers, it is a transit AS. If not, it is a stub AS. Nowadays, most of the ASes are multi-homed. They subscribe to more than one ISP service and have more than one link to connect with different ISPs to have a better service. On the other hand, a *single-homed AS* subscribes to only one ISP service [7]. Another type of ASes is stub. A stub AS does not have any customer ASes. If a stub AS subscribes to more than one ISP, similar to a leaf node in a tree with multiple parents, it is a multi-homed stub AS.

Multi-homing has become more popular as the cost of subscription to more than one ISP has decreased. Many ASes connect to more than one upstream AS in order to have a reliable service. We refer these connections as links. If one of the links is down, another link can act as a back-up link. These upstream links can share the network traffic load by applying traffic engineering. However, the multi-homed ASes need to control the inbound traffic. If one of the links is overloaded, this link is congested. Most of the time, we cannot control the inbound traffic easily because the routing paths are chosen by other ASes' routers.

Figure 2.1 illustrates a simplified version of the Internet. AS1 is a multihomed stub AS which connects to more than one upstream AS. AS2, AS3, AS4, AS5, and AS6 are transit ASes and they carry network traffic for other ASes. Each AS has interior routers and border routers for routing traffic. Interior routers are used within an AS or a network. Border routers are used for exchanging the reachability information with other ASes. They only use the Border Gateway Protocol (BGP), which is the only inter-domain routing protocol and is in its fourth version (BGP-4) [43]. After the border routers have received reachability information of the destination IP addresses, they decide the routing paths to different destinations. The routers pass their decisions to other routers within the same AS. If any node within an AS sends traffic to a destination IP address, the routers in this AS match the destination IP address with the reachability information and forward the traffic to the next-hop router according to routing path.

Reachability information for a given prefix originates from the AS which the prefix belongs to. It contains prefix path attributes allowing ASes to apply routing policies. For example, when an AS has connections to multiple ISPs (or multiple ASes), they send their reachability information of prefixes to other ASes according to their routing policies. This AS also receives reachability information from different ISPs. It can decide which information to use according to its own routing policy and the path attributes.

BGP allows ASes to choose their own administrative and routing poli-

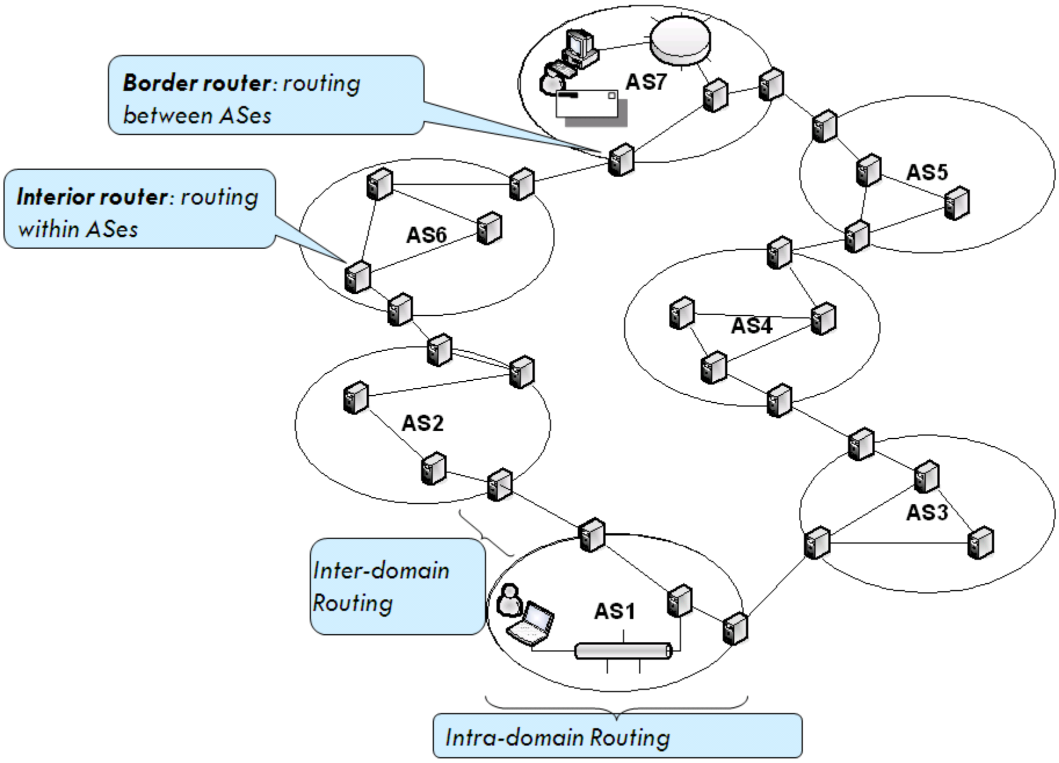

Border router: routing between ASes
Interior router: routing within ASes
AS7
AS5
AS6
AS4
AS2
AS3
AS1
Inter-domain Routing
Intra-domain Routing

Figure 2.1: An example of interdomain routing.

cies. The routing policies determine the route selection and propagation of the reachability information for each prefix to other ASes. When other ASes receive this information, they can selectively propagate it from AS to AS by means of BGP messages known as route updates.Routing information in BGP is route announcements. In this section, we will introduce BGP and routing policies in detail.

2.1.1 Border Gateway Protocol (BGP)

Border Gateway Protocol (BGP) [43] allows ASes to exchange reachability information through their border routers. There are different types of reachability information which include route announcement, withdrawal, and update.

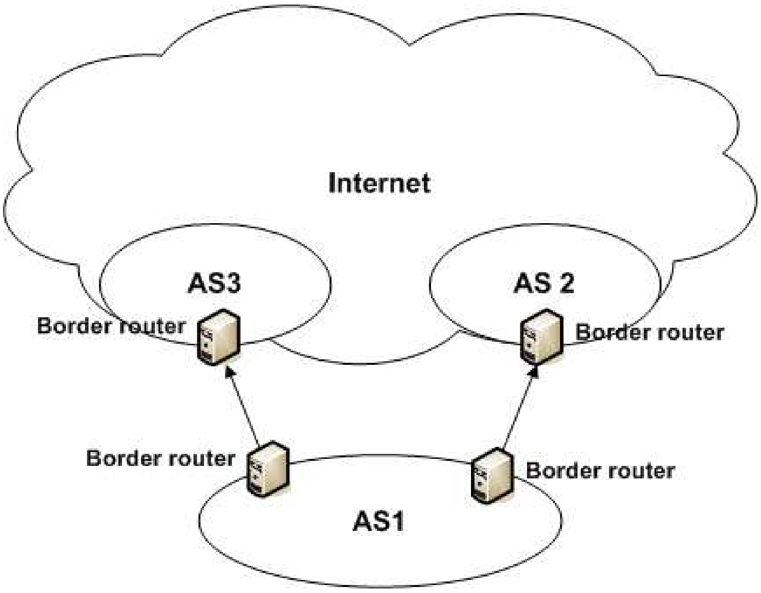


Figure 2.2: Multi-homed AS, AS1 and its upstream ASes, AS2 and AS3.

We can simply call these information route updates. An AS can announce a new prefix (perhaps a more specific prefix than the one already announced), withdraw an existing prefix, and update an existing prefix with different BGP path attributes. With the reachability information, a router can decide the routing path of different destination prefixes and send traffic to prefixes with the decided routing path.

In the process of forming a routing path, different routers are involved and send the reachability information to each other. When a border router of an upstream AS receives a route update of a prefix from its neighbor routers, the border router filters the routes, select the best route, and further propagate the information to other ASes. The border router first filters out the routes which are not considered in its routing policy. An import filter is used to filter out updates that should not be considered based on its routing policy.

After import filtering, the border router decides whether the newly received update is the best route. Each border router has a routing table which stores the best routes to each destination prefix. After it compares the new route to the entry of the prefix in its routing table, if the new route is better than the current entry, i.e. this new received route is the best route to the destination, it can replace the current entry of this prefix to the new one in its routing tables. This process is done by the route selection algorithm. BGP routers decide which route should be used in the routing table. When a BGP router receives more than one reachability information about the same prefix (more than one route is available to the same prefix), it will choose the best route according to the BGP route selection algorithm and place it into the forwarding table. At the same time, it may store those non-best routes for cases when the best route is not available.

After the router decided the best route of the destination prefix, the border router may further propagate this information to other upstream border routers. An output filter is used to decide which next-hop AS the information would be further sent to. Thus, when an AS announces a new prefix, or changes its input, or output filter, routers of the AS send out new updates to other routers and introduce routing dynamics.

In Figure 2.1, in order for other ASes to send traffic to prefix in AS1, AS1 sends out the route announcement to AS2 and AS3. The route announcement includes the prefix AS1 announced and BGP path attributes of the prefix. If the routers of AS2 and AS3 do not have the entries of the AS1 prefixes in their routing tables, their route selection algorithm selects the route announcement as the best route and puts it in their routing tables. After updating their own routing tables, if the routers in AS2 and AS3's output filters allow them to further propagate this information to other routers in AS6 and AS4, respectively, they would send out the update information to them. Every router performs the same mechanism, input filter, route selection algorithm, and output filter. Because routers within the same AS update each other about what they receive from other routers, routers in AS7 receive more than one route announcements of the prefixes in AS1, one of them through AS2, and the other through AS3. The routers in AS7 need to select which one is the best route and put it into their routing tables. If the routers choose the route through AS2 as the best route, they will put that route in their routing table.

With propagation of the reachability information, each router has routing paths to each prefix. The traffic of each prefix is sent according to the entries, which is the best route of each prefix, in the routing table and along routers on the routing path. The routing path is decided by the routers' decisions. If any node within an AS sends traffic to a destination prefix, the router forwards the traffic to the next-hop router according to the entry in the routing table. In the example (Figure 2.1), when a node in AS7 sends traffic to the prefix in AS1, the routers will forward the traffic through AS6, the routers learn the route from, and through AS2 to AS1. Thus, the routing updates change the routing paths and redistribute the traffic load on them.

2.1.1.1 BGP Routing Process

BGP routing process includes three steps, input filtering, select best routes by route selection algorithm, and output filtering [42]. In BGP, routers process and announce the reachability information of prefixes. The reachability information is called route announcements (or route advertisement). Each route announcement contains BGP path attributes of a prefix. First, for each BGP neighbor router, the administrator specifies an input filter that selects the acceptable route announcements. The selection criteria could be that an *AS Path* attribute, which is one of the path attributes, includes a set of trusted AS numbers. Or check if the next-hop IP address, another path attribute, is available. Second, when the input filter has accepted the route announcements, they are passed to the route selection algorithm. Once the algorithm selects the best route, BGP routers place it in the BGP routing table. Third, the router sends the best route to other routers (some of the attributes could be updated beforehand). After this operation, the BGP routing table contains all the acceptable routes received from the BGP neighbors.

The route selection algorithm is generally composed of seven criteria shown in Table 2.1. When a BGP router receives more than one reachability information, or *routes* for short, of a prefix, it selects the best route according to the sequence of these criteria in the route selection algorithm. If the first criterion cannot distinguish a best route, i.e. the local preference values of all the received routes about the same prefix are the same, the router will compare them based on the next criterion, i.e. their AS path length and so on. The first criterion, higher local preference, is based on the assigned local preference values of each route for each destination prefix. It facilitates the AS to prefer a certain border router, or link, to send traffic to. If the AS path length appears to be the same as the received routes, then the following criteria will be used to select the best route. We call them tie-breaking criteria. Finally, the route with the lowest BGP router ID will be chosen as the best route.

1. Higher local preference
2. Shorter AS path
3. Lower origin type
4. Lower MED value
5. E-BGP over I-BGP routes
6. Lower IGP metric to next-hop
7. Lower BGP router ID

Table 2.1: BGP route selection criteria

The propagation of the route announcement to other BGP neighbors involves the output filter. A BGP router uses the output filter to decide whether it should announce the best route in the BGP routing table to each BGP neighbor router. The BGP router can announce at most one route for each reachable destination to its neighbor routers.

2.2 Routing Policies

Each AS has its own routing policy that shows which route the AS would prefer to use and allow other ASes to use. The routing policy of an AS affects the traffic load of the ASes that are directly or indirectly connected with the AS. Usually, ASes apply their routing policies based on the relationships with their neighbors.

The routing policy includes import and export policies. The import policy contains the criteria that determine which routes are preferred to send traffic to the destination prefixes. The preferred routes are then sent to the route selection algorithm where the best route can be determined. After the best route is chosen, the router finally applies an export policy to select which neighbor the best route should be exported to. Then that neighbor can use this route to send traffic to the destination prefix.

2.2.1 Import Policy

The way to set up an import policy is through the use of an import filter and local preference values. An import filter can filter out the routes that are not preferred, and the routes that remain are passed to the route decision process. In this process the router selects the best route based on criteria such as local preference values. Usually the first route selection criterion is highest local preference values, which can influence the selection of the best route and control the outgoing network traffic. According to [49], one of the most important aspects of an import policy is the assignment of local preference values to each route.

To influence the preference of next-hop ASes, network operators assign the local preference values of each route for each prefix. If multiple next-hop ASes send route announcements of the same prefix to an AS, it will choose the one with the highest local preference. Then all the traffic going to that prefix will be sent through the next-hop AS associated with that announcement. Thus, network operators can control the traffic being sent to different next-hop ASes through the assignment of local preference values. The import policy can be applied in the following policies:

- Routes with the highest local preference. Gao and Wang [49] defined 2 types of routing policies onsetting local preference:

  Typical Local Preference : customer routes have higher local preference than the peer routes and then come to the provider routes.

  Atypical Local Preference : peer routes or provider routes have the same or higher local preference than the customer routes, or the provider routes have the same or higher local preference than the peer routes.

  In [49]'s finding, most of the prefixes have typical local preference for each AS.

In the typical local preference policy, network operators assign higher local preference values to the border routers connected to the customer ASes. On the other hand, in the atypical local preference policy, they assign the same value to all the routers regardless of customer, provider, or peer routes.

2.2.2 Export Policy

After running the route selection algorithm, a router follows the export policies of the AS it belongs to. It is done by an export filter. The export filter determines which best route it would advertise to the specified neighbor ASes. It gives these neighbor ASes a preference to use this route and send traffic through this AS which sent out the announcement. After a neighbor AS receives a route announcement, it can send traffic to the prefix in the announcement through the AS which sent out the route announcement earlier.

However, each AS has its own export filter. As a result, the inbound traffic of all the ASes on the path is affected. Export policies include permitting or denying a route, assigning MEDs to control the inbound traffic, tagging a BGP community to indicate what preference a neighboring AS should assign to a route, and prepending*AS Path*s or redistributing its prefixes to affect the inbound traffic. Some well known policies are [49]:

Exporting to provider: A customer can export its routes and the routes learned from its own customers to its providers, but cannot export routes learned from other providers or peers.

Exporting to customer: A provider can export its routes, the routes learned from the other customers, its providers and peers to its customers.

Exporting to peer: A peer can export its routes and the routes learned from its customers to another peer, but cannot export the routes learned from its providers and other peers.

With these policies, the customers only carry traffic for their customers. However, they do not carry traffic for the providers.

2.3 Inbound Traffic Engineering

There are limited methods to influence the inbound traffic on the routing level. Selective announcement, specific prefix advertisement, Multi-Exit Discriminator (MED), and AS path prepending can be used to control inbound traffic.

2.3.1 Selective Announcement and Specific Prefix Advertisement

An AS can selectively announce the prefix of an AS to different links[1] without overlapping, which is called selective announcement. It can also announce both the prefix and a more specific prefix of that prefix to one of the links, which is called specific prefix advertisement. Selective announcement can be used to distribute the traffic of different prefix to different links, but without any resilience support. If one of the links is broken, the network operator is required to announce all the prefixes to the un-broken links. On the other hand, in a specific prefix advertisement, an AS announces the shorter prefix to both links and more specific prefix to one of the links. Since routers prefer more specific prefixes, routers would choose the more specific prefix first and forward the traffic through the link to that prefix. As a result, most of the traffic is distributed according to the link of the specific prefix. However, most of the routers nowadays do not take prefixes longer than /24 [2]. This method is not very effective to further split the prefix. On the other hand, more routes are introduced to the routing table which increase the routing table size and the load of the routers.

[1] A link can be a physical link which connects the border router to the next-hop AS's border router.

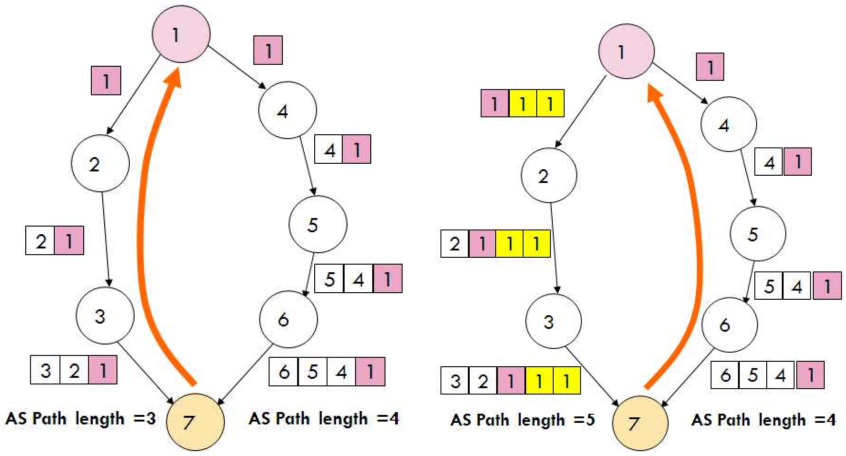
[2] xx indicates the subnetmask of the prefix.

2.3.2 Multi-Exit Discriminator (MED) for Inbound Traffic Control

MED is an optional, non-transitive attribute which will not be further transmitted to other neighbor AS. It tells an external neighbor about the preferred link to an AS while there is more than one link between them. An AS can assign different MED values to different links which connect to the same neighbor AS. A router attaches an MED value, which associates with one of the links, to the route announcement and sends the announcement to the external neighbor router. When the external neighbor router receives announcements from different links, it decides which link to use according to the MED value. In the route selection algorithm, lower MED value is one of the criteria. The router will choose the best route with the lowest MED value. As a result, the link with the lowest MED value is used to send traffic to the destination prefix. MED value is not used for inbound traffic engineering of multi-homing with multiple upstream ASes.

2.3.3 AS Path Prepending

In the route selection algorithm, AS path length is the next criteria below the higher local preference. Since local preference is controlled by the traffic source, not the destination prefix, AS path prepending is considered to be one of the prevalent methods to control inbound traffic. An AS prepends its own AS number more than one time on an AS path of the announcement to one of the upstream links, so the AS path through that upstream link is longer



(a) Before performing AS path prepending

(b) After performing AS path prepending with length 2

Figure 2.3: AS1 announces its prefix to both upstream ASes, AS2 and AS4. AS7 received two route reachability information of that prefix. It chooses the best route according to the route selection algorithm in BGP and its routing policies.

after prepending. Thus, prepending prevents other ASes to choose this link as a best route. The AS propagates the reachability information further to next-hop AS. When an AS receives more than one route, it compares the AS path lengths of them if the local preference values are the same.

Figure 2.3 illustrates how prepending works. Each node (circle) is an AS. AS1 announces its prefix to both upstream ASes, AS2 and AS4. After the route announcement propagates to other ASes and arrives to AS7, AS7 received two route announcements of the prefix of AS1. AS7 then decides which route should be used according to the route selection algorithm introduced in Section 2.1.1.1 (Figure 2.3(a)). If AS7 does not have a higher local preference to any upstream ASes, AS3 or AS6, then it chooses the best route by the shorter AS path. The routing path on the left hand side is shorter and thus, AS7 chooses this routing path and places it in its forwarding table. All the routers in AS7 send traffic to that prefix according to the entry in the forwarding table. It may also further propagate this information to other ASes.

When AS1 wants to shift some of the incoming traffic from AS2 to AS4, it prepends its own AS number more than once on the AS path when it announces the prefix to AS2 while keeping the announcement to AS4 unchanged (Figure 2.3(b)). After these route announcements reach AS7, because the AS path length on the left side is longer than that of the right side, AS7 decides to change and send traffic to the prefix through AS6. As a result, AS1 can shift some of the incoming traffic to another non-prepended route.

AS path prepending provides network resilience. When there is any link failure and the BGP session between routers fails, routers immediately look for other possible routes to all the prefixes in the forwarding tables. As a result, if an AS uses AS path prepending to avoid traffic from one of the upstream and there is a link failure on the preferred link, the non-preferred link is still available for the routers on the remote ASes to switch to.

However, the effect of prepending is unpredictable. The topology and routing policies of ASes are unknown. Some of the ASes have already set a higher local preference to some of the outbound links such that we cannot influence them with prepending. Furthermore, when the prepending comes up with routes of the same path lengths, tie-breaking is considered. Both routing policies and topology of upstream ASes affect the result of prepending. Network operators often use it in a trial-and-error basis which may introduce a large amount of network churn.

2.4 Studying the Routing Dynamics by Passive Measurement

Generally, we can study the routing dynamics by performing measurements. Measurements can be categorized to passive and active measurement. Passive measurement uses the publicly available BGP routing data archives collected from vantage points to study the routing dynamics. We define active measurement as we actively introduce routing dynamics by injecting a route announcement to the Internet and observe the results.

In passive measurement, we examine measurement results previously derived from Route-Views's data to discuss what we can and cannot infer from them [51]. Route-Views has started getting routing tables from several vantage points since November, 1997. Now, 94 ASes are currently contributing views to Route-Views collectors. These ASes include: Level3, Sprint, AT&T, UUNET, etc.

In this section, we introduce some of the results from passive measurement. The results include multi-homing phenomenon, prepending, and the policies ASes used to apply prepending. These statistics tell us that prepending is one of the prevalence methods nowadays and research on it is important.

2.4.1 The Growth of the Multi-homing Phenomenon

The trends on the numbers of ASes, stub ASes, multi-homed stub ASes, and transit ASes are shown in Figure 2.4. The increasing trend of multi-homed stub ASes gives us the motivation to research traffic engineering for multi-homed stub ASes. The number of stub ASes is around 80% of the total number of ASes. The percentage of stub ASes connected with more than one upstream AS increased from 40% in Nov, 1997 to 60% Jun, 2006. The data shows that one third of the multi-homed stub ASes have been implemented with prepending. 16% of the ASes are transit ASes. The transit ASes' routing policies are determined by commercial contracts, so their traffic engineering is done according to the commercial contracts as well. Thus, transit ASes can perform traffic engineering in their own way. However, the multi-homed stub ASes need to determine the most effect method to control traffic. This motivates us to study the inbound traffic engineering for multi-homed stub ASes. Some dips in Figure 2.4 have appeared. This is because of various worm attacks or blackouts. For example, for the August 15, 2003 statistics, there is a dip at that time. It is the exact day of the blackout. In [18], some ISPs and banks were affected at that time.

2.4.2 Prepended Routes Observed in the Route-Views Archives

A prepended route is one that contains duplicate AS numbers on its AS path. These duplicate AS numbers may have been inserted by the origin AS, by

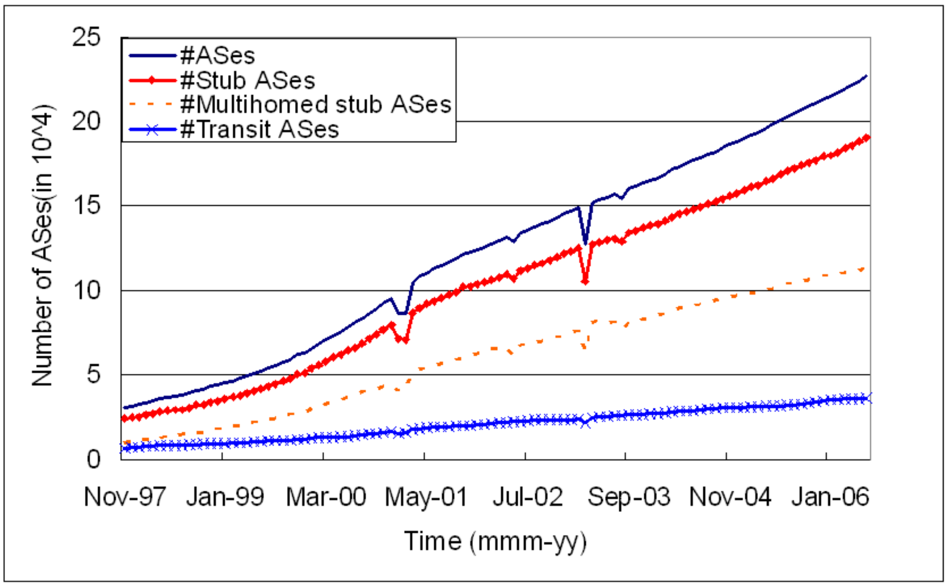


Figure 2.4: The trends on the numbers of stub ASes, multi-homed stub ASes, and transit ASes.

another AS on the AS path, or both. Figure 2.5 shows the total number of routes and the number of prepended routes observed by the Route-Views's collectors over time. Various factors are responsible for the increase in the number of routes. The most obvious one is the increase in the number of ASes in the Internet throughout 1997 to 2006. Another is the increase in the number of Route-Views peers. The new peers provide different views of the network, and thus, more routes can be observed. Moreover, many prefixes are split into multiple prefixes as a result of selective announcements and traffic engineering[9]. Each route is responsible for one prefix. As can be seen in Figure 2.5, the ratio of prepended routes to total routes has remained fairly constant over time. However, this statistic cannot reflect the exact situation of using prepending to perform inbound traffic engineering. We will further talk about this in later sections.

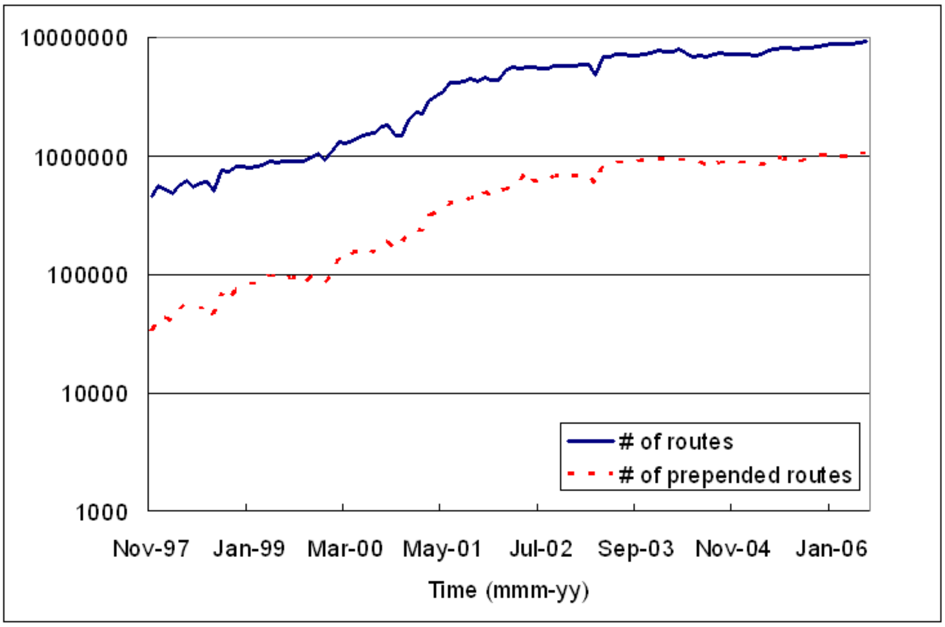


Figure 2.5: The number of prepended routes observed in the RouteViews archive.

2.4.3 Prepending Policies

We refer the ways to implement prepending as prepending policies. We study the characteristics of the route announcements and classify them into different ways of applying prepending. The prepending characteristics of a route can be classified into source prepending and intermediary prepending. Source prepending refers to the prepending which is performed by the AS who originates the route, while the intermediary prepending is performed by the transit ASes of the route. For example, the AS path ($AS6$ $AS2$ $AS2$ $AS1$ $AS1$ $AS1$) is originated by AS1. AS6 announces this route which indicates that AS6 can reach AS1 through AS2. This route has an AS path length of 6. AS1 is said to perform source prepending and AS2 is said to perform intermediary prepending. As shown in this example, a route could have both types of prepending or more than one intermediary prepending.

In Figure 2.6 70% and 30% of the prepended routes have source prepending or intermediary prepending, respectively. About 4% of the prepended routes have both source and intermediary prepending. Note that the actual percentages for the routes having both prependings would be higher because the routers filter out the routes with longer AS path and these routes were not further propagated to the Route-Views router.

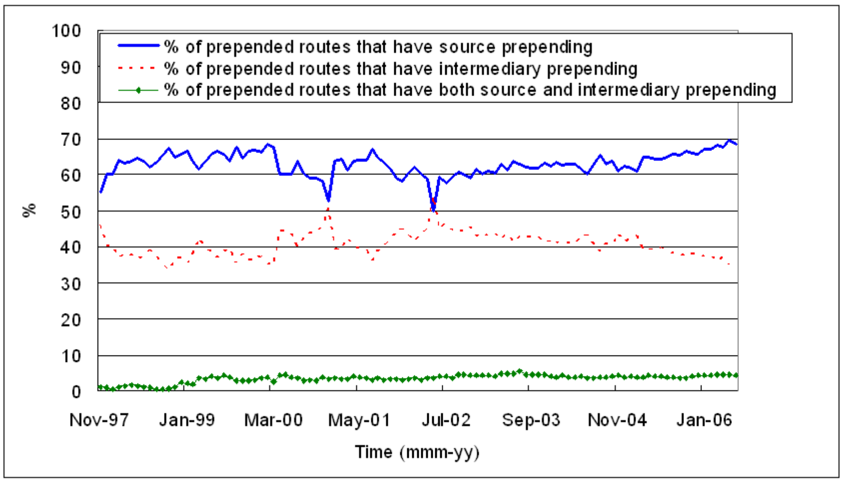


Figure 2.6: The percentages of source prepending, intermediary prepending, and mixed prepending in the prepended routes.

Another characteristic of the prepended routes is the distribution of the number of prependings in a route and the distribution of the prepending length. In the former example of ($AS6$ $AS2$ $AS2$ $AS1$ $AS1$ $AS1$), there are 2 prependings in this route. The prepending length of the prepending done by AS2 is 1, while the AS1's prepending has a length of 2. The distribution of the number of prependings in a route and the distribution of the prepending length are shown in the Figure 2.8 and Figure 2.9, respectively.

Figure 2.8 shows that around 95% of the prepended routes have only one prepending and around 4% of the prepended routes have 2 prependings.

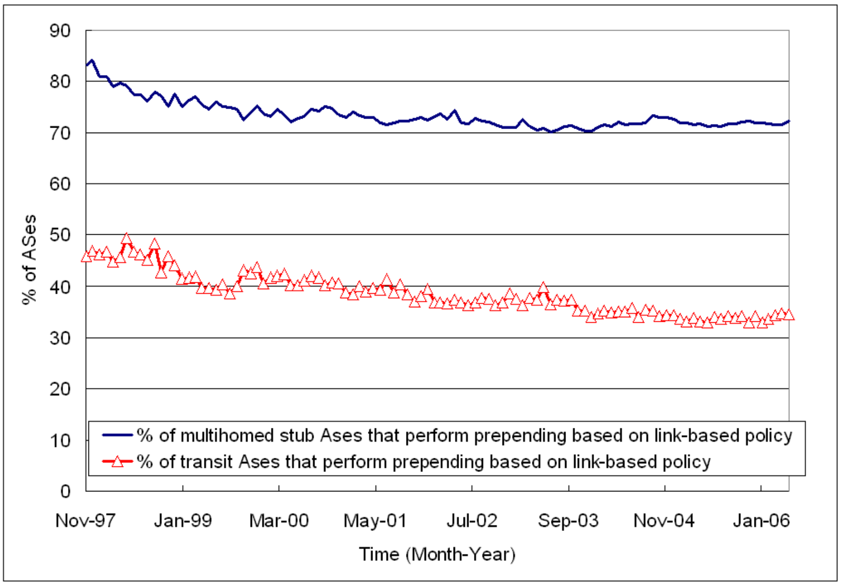


Figure 2.7: The percentages of ASes that employ the link-based prepending policy.

Furthermore, Figure 2.9 shows that the prepending length of 1 is the most common case we found. Some of the percentage of prepending length of 1 is shifted to prepending length of 2 since 2003. Since 2003, there has been an increasing percentage of prependings with prepending length of 2 and a decreasing percentage of prependings with prepending length of 1. Possible reasons for this shift are that routers may ignore the routes with prependings shorter than 2, or they cannot find routes better than those prependings with prepending length of 2. Normally routers choose routes with shorter AS paths. If we observe more routes with prepending length of 2 or above, fewer routes with prepending length less than 2 exist for the routers to choose. Another reason is that the ASes performed prepending more frequently. Thus, they require prepending with longer length in order to be effective.

The prepending policies can be classified into more meaningful types, such as link-based prepending and prefix-based prepending. If an AS performed

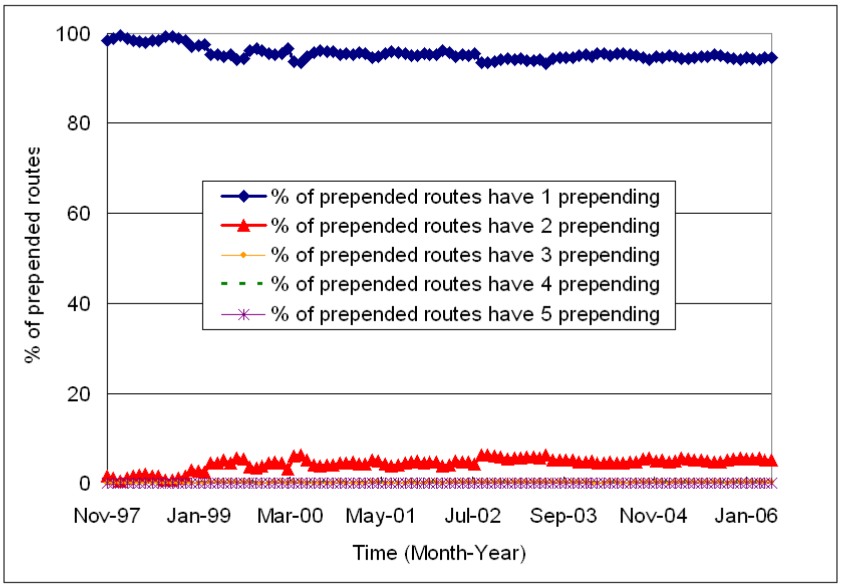


Figure 2.8: The percentages of prepended routes that have one or more prepending.

prepending on all its prefixes with the same prepending length as one of its upstream ASes, the link is said to be employed with a link-based prepending policy. Otherwise, the link is said to be employed with prefix-based prepending policy if the prepending lengths of the prefixes announced to this link or upstream AS are different across this link.

Figure 2.7 shows the statistics of the percentages of ASes that employed the link-based prepending policy. The percentages of the multi-homed stub and transit ASes and their links that performed link-based prepending policy are decreasing. One of the possible reasons is that the ASes implement more complicated prepending policies than link-based prepending policies. Thus, they prepend the prefixes differently on each link.

However, we note that due to BGP's nature, this data from passive measurement is not sufficient to quantify the effectiveness of prepending. The purpose of prepending is to discourage using a given route. Therefore, if the prepending is successful, the prepended route will not be selected as a best route and not be preferred. Since only best routes are announced, the non-preferred prepended routes will not be propagated and it will not appear in the Route-Views data. Thus, we cannot measure how effective prepending is simply based on counting prepended routes observed in passive measurement data. The more effective it is, the fewer times we will observe it.

To probe further, in Figure 2.9 we show the distribution of prepending lengths over time. We refer to the prepended routes with a prepended length of $i \geq 1$ as i-prepended routes. That is, the AS path attribute contains $i + 1$ identical AS numbers. At least 50% are 1-prepended routes, followed by 2-prepended routes and 3-prepended routes. The most noticeable change in recent years is the increase in the share of prepending lengths of four or more.

The presence of a relatively large number of 1-prepended routes is probably due to the fact that a prepending length of one is not sufficient to affect the original routes. As a result, these 1-prepended routes are still preferred and selected as best routes. This can apply to other lengths as well. However, a longer prepending length is less likely to be preferred. Therefore, we should see fewer visible routes with high prepending lengths. As for the increased share for $i \geq 4$, one possible explanation could be that some operators increased the prepending length beyond three, and afterwards discovered that a shorter length is not sufficient to affect traffic. Once again, passive measurement data alone cannot help us understand whether a longer prepending length is effective or not. As we shall see in the following section, the use of active measurements can overcome this problem.

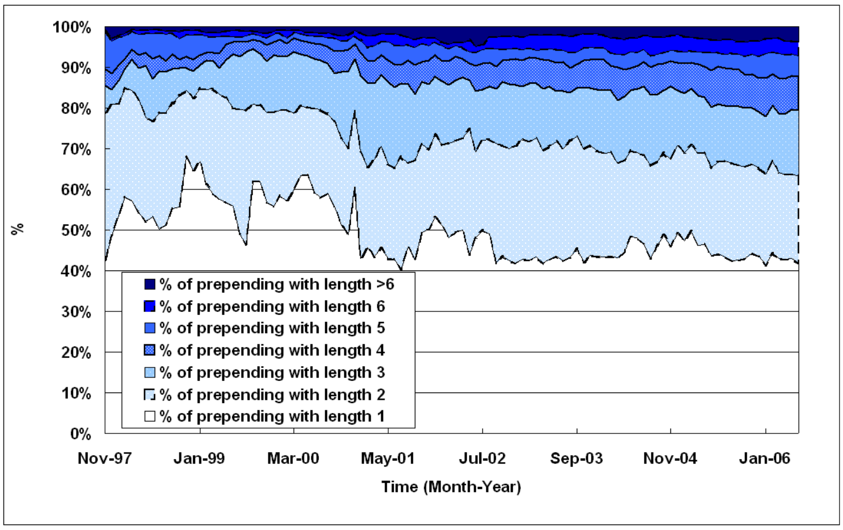


Figure 2.9: The distribution of prepending lengths observed in the Route-Views archive.

2.5 Available Resources for Studying the Routing Dynamics

Vantage points (VPs) are available for us to study the Internet routing activities. VPs are machines which are available to the public to retrieve the BGP routing information of any prefixes based on the location of the VPs in the Internet. VPs tell users about the routing paths available to the ASes where the VPs are located. Thus, we can check the best routes or available routes to those ASes through VPs.

2.5.1 Types of Vantage Points (VPs)

Vantage Points (VPs) provide the routing information in the form of routing tables and archive the tables into data files. VPs can also provide the routing information in the form of BGP route updates. These have the timestamps of the route updates which gives more detail on how the BGP updates triggered and propagated.

VPs can be looking glasses (LGs) or route servers (RSs). LG provides `show ip bgp` command on a web interface. A user can parse a prefix into the query through a LG and retrieve the prefix's routing path available to the AS which the LG located in. A RS is a telnet machine and is also used to observe the routing changes. We can use the BGP commands to retrieve the routing information of the prefix we want. We can also review the BGP routing table of the AS which the RS is located in. At the same time, the BGP routing table reveals all the routes available to the AS from the its upstream ASes or peers.

We can also use PlanetLab [39] as a source of VPs. PlanetLab is a global research network that provides facilities for the development of new network services. It has 808 nodes and allows users to deploy their network services on these nodes. We can use these nodes to perform traceroute [33] to the IP addresses we are studying. IP traceroute gives IP routing paths from the PlanetLab nodes to the IP address we are querying. Then we translate the IP addresses on the IP path into AS numbers to obtain the AS routing path. However, this is not the same as the BGP AS paths since the IP-AS path does not include any prepending information, so we cannot recognize the AS number of some IP addresses, e.g. private IP addresses on some routers in the internet exchange etc. [35].

2.5.2 Route-Views and RIPE RIS Routing Data Archives

Route-Views [4] and RIPE RIS [44] provide routing data archives for the public. They peer to ASes in different locations in the Internet. Route-Views collects the routes from its peers (around 50 peers). After it receives route updates from these peers, it updates its routing table. It also archives its routing table every two hours along with BGP updates in MRT format. The BGP update archives require [45] to decode them. Route-Views also has route servers for the public to query it by using `show ip bgp PREFIX` command.

RIPE Network Coordination Center (RIPE NCC) is one of the Regional Internet Registries (RIRs) [37] providing Internet resource allocations, registration services, and coordination activities for the Internet operation. It has 16 remote route collectors(RRCs) located in Europe and North America to collect Internet routing and performance information from its members. Nowadays, only 14 of them are active. RIPE Routing Information Service (RIS) makes this information available to the community for troubleshooting and research. Each RRC collects BGP routing information from its peers. RIPE archives the BGP routing tables from each RRCs every 8 hours and BGP updates every 15 minutes. RIPE also provides looking glasses web interface for the public to retrieve the information of certain prefixes.

2.6 Related Works

AS path prepending is one of the inbound traffic engineering methods. RFC3272 [5] described the principles of traffic engineering. It discussed architectures and methodologies for performance evaluation and optimization of operational IP networks. In this section, we introduced related works of different approaches to study AS path prepending and inbound traffic engineering.

2.6.1 Passive Approaches

In earlier studies, Feamster et al. [21] and Broido et al. [8] reported their findings on prepending based on the AT&T backbone data and Route-Views data, respectively. In a recent passive measurement study, Wang et al. [51] gave a more up-to-date report using various statistics compiled from the Route-Views data and proposed a model to study fundamental issues of decentralized traffic engineering in the Internet.

2.6.2 Active Approaches

In terms of active approaches, Chang and Lo [12] proposed `AutoPrepend`, an automated procedure that determines the best prepending length before affecting the change. A key component of `AutoPrepend` is the use of a beacon prefix to predict the magnitude of the incoming traffic volume variation due to prepending. Recently, Quoitin et al. [41] conducted similar active measurement experiments to study the prepending method, and evaluated the effect on inbound traffic distribution on different incoming links. Moreover, they built a new simulator to study the prepending method, and reported a number of findings.

However, neither of these works attempts to analyze the AS-level route changes induced by prepending. On the other hand, the goal of our active measurement is to understand the effects of prepending on AS-level routes. To archive this, we observe route changes induced by prepending from a number of vantage points in the Internet. Although [41] also examines routing tables in some upstream ASes, that methodology is practically infeasible to achieve the scope of measurement described in this dissertation. By observing and analyzing route changes, we expect to be able to explain some of the results observed in [12] and [41]. Furthermore, the analysis of the results obtained using our active measurement methodology could be used to improve `AutoPrepend`'s prediction accuracy.

Finally, we note that prior to this work, we conducted a set of preliminary active experiments for a stub AS [32].

2.6.3 Optimization of Prepending Length

Recent work has focused on the development of methodologies to determine optimal prepending lengths. Gao et al. [22] and Di Battista et al. [6] have proposed algorithms to solve optimization problems for the prepending method. These methods assume that all routers in an AS select the same route and make routing decisions based only on the shortest AS path. However, these assumptions do not hold in practice. As we can see in §chapter 5, our results show that this is often not the case.

Chapter 3

Active Measurement

Our active measurement methodology can facilitate network operators and researchers to measure routing dynamics induced by any network operating activities on BGP. The measurement is ideally designed to study AS path prepending, which is one of the inbound traffic engineering methods. When ASes receive a prepended route, their routers can accept or decline this path as a best route. The active measurement allows network operators to observe which path the ASes have chosen as best path and further study the topology and other factors of their choices.

It is difficult to predict the effect of network operating activities. First, when an AS announces a route, it cannot control how other ASes will choose the routing path. It is because they have their own routing policies and can send traffic through different paths. Second, some of the ASes are single-homed. These single-homed ASes have a limited choice of routes, and can choose the best routes only from their providers. Thus, their effects to any network operating activity are based on their providers. Third, different ASes have different amounts of traffic to send to their destinations depending on the applications they hold. ASes can adjust the routes according to their traffic load.

We have three components in our measurement infrastructure: beacon and control prefixes, route announcements, and vantage points. We announce a beacon prefix to the selected link(s) with prepending and to other links without prepending, and announce control prefixes without any prepending to all the links. We define a *link* as a BGP session between two adjacent ASes. In our case, a link is between the AS which announces prepended routes and its upstream providers. Then we collect the resulting routes from a set of well-spread locations called vantage points, and compare the resulting routes to the routes seen in the absence of prepending.

The setup of this measurement is simple, flexible, and will not generate traffic to disturb the production network with the exception of traffic of BGP announcements. We can use the existing facilities, e.g. border routers or software routers, to actively announce the beacon and control prefixes. The network operators can study some of the upstream providers or all of them by announcing the prefixes to some of the upstream providers. For observing the routing dynamics, existing tools facilitate the measurement of the announced prefixes. Existing tools include looking glasses, which support BGP queries or traceroute, and route servers, which support BGP queries. These tools are separate from the route announcement. As a result, any network can apply these measurements to evaluate the routing dynamics induced by prepending in other ASes.

3.1 Key Components

We describe the three key components in the active measurement methodology: beacon and control prefixes, route announcer, and vantage points.

3.1.1 Beacon and Control Prefixes

To minimize disruption of normal Internet traffic, we send BGP updates of a set of *beacon prefixes* instead of the operational prefixes. The beacon prefixes must be shorter than `/24` in order to prevent upstream routers from filtering them out, or aggregating them with some shorter prefixes. The rest of the Internet includes different upstream ASes and treats the beacon prefixes the same way as prefixes used for production traffic. As a result, the effects of prepending observed on the beacon prefixes are representative of the behavior of other prefixes for production traffic.

To ensure that path changes observed are induced by our active measurement, and not other network events or topology changes, we announce a number of *control prefixes* without any prepending on it. The control prefixes give information about any activity on the Internet that is not under prepending. When we observe any event on the beacon prefix and control prefix at the same time, that event possibly does not relate to the prepending on the beacon prefix and we should not take it into our measurement.

3.1.2 Route Announcements

We announce the beacon prefixes with different prepending lengths or any other type of announcement we want to measure. All the route announcements require special arrangement with network operators and upstream providers, just like prepending in operational networks. We first schedule the announcements without interrupting normal router operations, i.e. it does not bring any failure to the normal BGP session connections which are announcing normal production prefixes. We also need to coordinate with upstream providers about the prepended route announcement such that they will not filter it. Moreover, the time between two consecutive announcements should be long enough for the newly announced routes to converge before measurement and for route flap damping [34] to expire. Furthermore, we announce control prefixes without any prepending, i.e. prepending length of zero, to all upstream providers at the same time as a control experiment.

3.1.3 Vantage Points

After we announce beacon and control prefixes, we want to collect the routing dynamics induced by the prepended route announcement. The routing dynamic induced by prepending show how other ASes choose the best route of these prefixes. Furthermore, we want to observe how the route announcements propagate to other ASes in the Internet.

To measure the routing dynamics, we use some publicly accessible sources of BGP information, or *vantage point*(VPs). VPs are located in different ASes for users to look at a network's routing information. If there are any route changes in a network, we can observe them through the network's vantage point. The main advantage of using VPs is that we can access them without any pre-arranged coordination with other ASes. A similar concept has been used for route convergence studies [34]. Common VPs are route servers, looking glasses, and reverse traceroute facilities. They provide either an AS path or IP path. Our measurement uses AS paths to analyze the effects of prepending because BGP evaluates AS path length instead of the IP path length.

The first two types of VPs, route servers and looking glasses, provide their own BGP routes or BGP routes learned from other peers. BGP routes contain information of AS path to different prefixes. Since route servers are telnet servers, users can use BGP router commands to access the routing information of the network it belongs to. A looking glass is a webpage or a router with a web interface which allows users to check the routing information of the network it belongs to.

On the other hand, a reverse traceroute requires a machine in a network. It can further be implemented with PlanetLab[39] nodes. Users can access that node and use traceroute commands to check the IP path between the network of the node and an IP address. IP paths can be converted to AS paths by IP-to-AS mapping.

Some of the networks collect routes from different peers and publish the routing information in their route servers. Router collectors are used in these networks.

- **Route Collectors**

  Route collectors are already set up for collecting routing information from peers. The route collectors we chose are publicly available to all users.

  - **Route-Views** University of Oregon Route Views Project [4] provides a tool for Internet operators to obtain real-time information about the global routing system from different backbones and locations around the Internet. It is equipped with both looking glasses and route servers for users to query the routing information. The routing information includes more than 6 peering points, each having more than 50 peers. The routing information is archived in both BGP updates and routing table snapshot format. The BGP updates are archived every 15 minutes while the routing table snapshot is taken every 2 hours. Some of the archives are in MRT format which require[45] to decode them.
  - **RIPE NCC RIS Remote Route Collectors**

    RIPE Network Coordination Center (RIPE NCC) is one of the Regional Internet Registries (RIRs) [46] providing Internet resource allocations, registration services and co-ordination activities for the Internet operation. It has 14 remote route collectors (RRCs) located in Europe and North America to collect Internet routing and performance Information from its members. At the same time, it makes the routing information available to community for troubleshooting and research. RIPE archives the BGP routing tables from each RRCs every 8 hours and BGP updates every 5 minutes. RIPE also provides looking glass web interfaces for public to retrieve the routing information immediately.

- **Looking Glasses** Looking glasses are web pages which allow users to query the routing tables of a router. Both Route-Views and RIPE have looking glasses for users to query their routing information. Many ASes provide publicly available looking glasses for users to query their routing information immediately. Traceroute.org introduces a set of looking glasses. However, not all of them provide BGP routing information. They may provide only traceroute facilities from the router to the IP address entered.

- **Traceroute** Traceroute [33] is a tool to diagnose the routing path on the IP level. It returns an IP address path between the source node to the destination node. ICMP [40] is used to retrieve this information, but for security reasons not all the nodes on the path enable this option. At the same time, a traceroute path is an IP path instead of an actual BGP AS path. The IP path must be converted to AS path in order to analyze the results on the BGP level[35].

  - **traceroute.org**

    traceroute.org [29] provides a lot of links of looking glasses which have traceroute utility. We can query those looking glasses to retrieve the reverse traceroute from that looking glass to a certain destination.

  - **traceroute.org** Traceroute.org [29] provides many links of looking glasses which have traceroute utility. We can query those looking glasses to retrieve the IP paths from these looking glasses to a certain IP address.

∗ **PlanetLab** PlanetLab consists of more than 800 nodes in the Internet for global research on development of new network services. These nodes are machines installed in different academic institutions and industrial research labs. We can login to any of these nodes and perform traceroutes to retrieve the IP paths from the node to a destination IP address. Since it does not provide BGP facilities for users to retrieve routing information of routers, we need to convert the IP paths into AS paths.

The data we collected from the vantage points are used to do the analysis. We compare whether there is any route change on both control and beacon prefixes. Because we want to observe the route changes induced by the prepended beacon prefixes, we only want to observe the changes induced by the beacon prefixes. The control prefixes are used to prevent observing route changes not induced by prepending. If there is any route changes observed on both beacon and control prefixes, the changes should be neglected. It is because these route changes are highly possible that they are not induced by prepending on the beacon prefix. A similar concept has been used for route convergence studies [34].

3.2 Procedures

The basic idea of the active measurement is to test the response of other ASes to different prepending lengths of beacon prefixes.

1. Announce a beacon prefix with a certain prepending length. At the same time, announce a control prefix without any prepending.

2. After at least two hours, which is the maximum time suggested by [53] for the network to converge, collect the route changes of these prefixes from those vantage points.

   The time between route announcement and route change collection affects the measurement results. If we do not wait long enough for the network to converge, the measurement results will not be accurate. As we mentioned, the maximum time for the network to converge is two hours [53] because of route flap damping, which is a BGP mechanism used to prevent frequent route updates. The two hour time to convergence is validated by simulations [53] of announcements and withdrawals in different sizes of networks.

3. Repeat step 1 and 2 with different prepending lengths on the beacon prefix.

After the route collection, we can further increase or decrease the prepending length and repeat the route collection. In the end, we can analyze the effects of different prepending lengths and observe the route changes. We will illustrate the measurement setup and implementation in the next chapter and the analysis in Chapter5.

3.3 Assumptions and Constraints

We have several assumptions and constraints. We assume that the beacon prefix is treated the same way as other prefixes of the same AS. We also have not taken the traffic volume into account. Also, VP distribution on the Internet and responses of the VPs are constraints to our measurement.

This active measurement methodology first assumes that other ASes treat the beacon and control prefixes the same as the other operational prefixes from the same AS. Thus, we can use the measurement results from the beacon prefix to observe the effects of prepending on other operational prefixes. However, when some of the ASes apply outbound traffic engineering. They pick some of the prefixes and set their routing path preferences, i.e. the exit points of the traffic to those prefixes. Although some ASes adjust their outbound traffic based on destination prefixes instead of ASes, we can still use the results to study the network. Because we announce the new beacon and control prefixes, these prefixes are treated normally and are most likely to reflect the effects of prepending on the existing prefixes.

3.3.1 Traffic Issues

We have not taken traffic volume into account in our measurement because our measurement is focused on the routing dynamics induced by prepending. To further observe the traffic volume shift, we can use NetFlow [16] to get the traffic volume of each prefix.

This measurement introduces two kinds of traffic: traffic from route announcements and traffic from retrieving routing information from VPs. Because this approach involves very little traffic, the disturbance to the operating network is minimized. The beacon and control prefixes do not carry any production traffic.

3.3.2 Route Filtering in Upstream ASes

Some upstream ASes filter the route announcements based on their AS paths [38]. These upstream ASes do not accept any announcements with prepending on their AS paths. If there is any prepending on the AS path, it is possible that the announcement is filtered by these upstream ASes. Thus, when we use prepending to perform inbound trafficengineering, we should coordinate with our immediate upstream providers and ensure that they do not filter out the prepended routes. Moreover, filtering by higher level upstream ASes is out of our control. There are some upstream ASes of our immediate upstream providers that may filter the prepended routes. However, if these filtering events have applied to our beacon and control prefixes, most likely they will also apply to our operational prefixes. As a result, we can measure the real responses of other ASes to the prependings of any prefixes within the AS which performs the measurement.

3.3.3 Locations of VPs

The number and the locations of VPs also affect the measurement results. VPs include Route-Views route servers, and looking glasses which provide BGP routing information. Before the measurement, we should study the location of VPs in order to retrieve a useful sample of results, even though the locations of the VPs are out of our control. At the same time, every AS has its own preference of path usage. For example, they may require better performance with networks in China rather than those in the US. In this case, VPs from China are more important. To solve this problem, we can replace the VPs by PlanetLab nodes [39]. PlanetLab nodes have a traceroute tool and have a more diverse geographic location. Traceroute gives us an IP path which we can convert into AS path. However, AS-level traceroute is still not entirely accurate [35]. If full accuracy in the AS path tracing is not required, the active measurement methodology can easily include PlanetLab nodes as additional VPs.

3.4 Possible Implementations

The active measurement is flexible and simple with a number of possible implementations. The route announcements can be performed by network operators who type in the command on the BGP border routers. We can also use software routers to perform announcements. Scripting can automate the whole route announcement and route collection processes. We can use publicly available VPs for route collection without any arrangement with them.

We have implemented the active measurement in two ways:

1. Route announcements were performed by network operators. Reverse traceroute and route servers were used to collect route updates.

2. Scripts were used to automate the route announcement with a software router. Route collection was done by collecting routes fromroute servers and looking glasses.

Chapter 4

Implementations of Active Measurement

Active measurement allows ASes to study and understand the effects of prepending to their ASes. Network operators can further use the data collected from the measurement to estimate the effects of prepending on other prefixes of their networks.

To study the routing dynamics of prepending on different networks, we have implemented the active measurement on two different ASes: a Hong Kong local university network which uses AS path prepending to control its inbound traffic, and RIPE Network Coordination Centre (NCC) [44] which is one of the Regional Internet Registries (RIRs) [37] providing global Internet resources and related services (IPv4, IPv6 and AS Number resources) to members in the RIPE NCC service region. Then we collect the route changes from different vantage points (VPs). The collection is done by a Linux machine equipped with scripts.

4.1 Implementation on a Hong Kong Local University Network

We have set up an active measurement facility in a Hong Kong local university network, which we simply call it *Home AS* in order to conceal its identity. We will call this implementation as "a local university network" for simplicity. This *Home AS* is a *stub AS* which does not transit any traffic for other ASes. *Home AS* has 2 upstream providers. We call the connection between each provider and *Home AS* as a link. We can observe the impact of the prepending on link in this implementation.

The route announcement infrastructure is comparatively simpler than the other implementation on RIPE NCC. As depicted in Figure 4.1, two BGP border routers are connected to AS1 (a tier-1 ISP) and AS2 (a regional ISP) respectively. These two BGP routers announce the BGP updates for a beacon prefix to AS1 and AS2. The beacon prefix is a set of addresses with prefix /21 in Home AS. Thus, the prefix is shorter than /24 and will not be filtered by upstream ASes due to the longer prefix length. At the same time, this beacon prefix was not being used. In other words, *Home AS* normally does not expect to receive traffic destined to the beacon prefix. Our measurement did not affect the normal operations of Home AS.

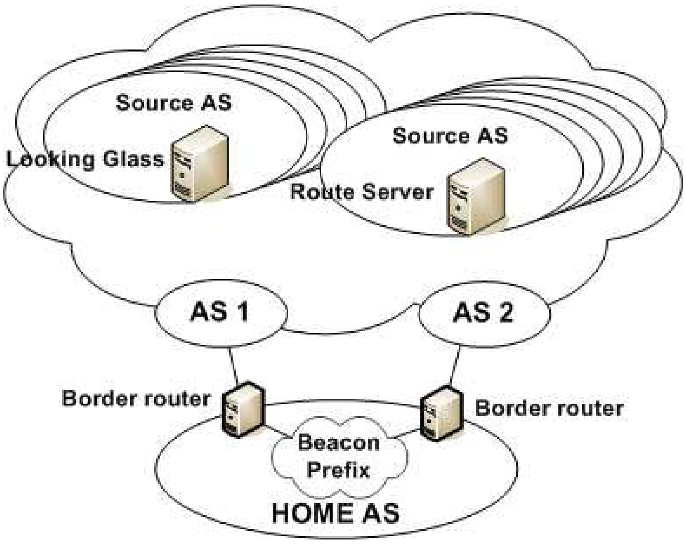


Figure 4.1: The active experiment setup at the Home AS.

4.1.1 Route Announcement Infrastructure

In the measurement, we announce the beacon prefix at the border routers with BGP router commands. We first notify the network operators of our upstream providers. They configure their routers to not filter the prepended routes on the beacon prefix. After that, we observe the route change at the looking glasses to ensure that the announcement is successful and does not affect other operating prefixes.

We announce the beacon prefix to all the upstream providers, but with prepending only on AS1. Because it takes time for route convergence in the Internet, we wait for at least five hours after the announcement [30] to collect the route changes from looking glasses and route servers. After we collect the route changes, we change the prepending length on AS1 and repeat the route collection. The overall objective is to study the impact of the prepending on AS1 on the routing paths for the traffic sources to reach Home AS prefixes.

We apply prepending on AS1 only because there is more incoming traffic through AS1, and Home AS applies prepending of other prefixes on AS1 as well. Essentially, the network operator increases the prepending length once, and then observes the change of the inbound traffic on both links. If the traffic via AS1 does not decrease much, the network operator further increases the prepending length on AS1. If it shifts too much traffic to AS2, the network operator decreases the prepending length on AS1 by one, and usually he can get the optimal length he/she wants. However, this trial-and-error based inbound traffic engineering would shift more traffic than expected, and induce congestion on another upstream AS. Furthermore, we have performed *forward prepending* by increasing the prepending length one by one from zero (i.e. no prepending) to five, and *backward prepending* by decreasing the prepending length to zero. The maximum prepending length is five, because we do not observe further route changes beyond a prepending length of four. We will discuss this in Chapter 5.

4.1.2 Route Change Collection Infrastructure

In this implementation, we use a set of vantage points (VPs), which include 16 route servers and 42 looking glasses. This will serve as a set of (virtual) traffic source and to allow us to collect the routing paths from these VPs to the beacon prefix. The route servers are located two to four AS hops away from `Home AS`. In contrast, the looking glasses are much farther away. Most of them are five to seven AS hops away. Therefore, we could also evaluate the impact of prepending with respect to the AS path length.

Route servers are BGP routers, or machines equipped with software routers. ASes set up route servers for their customers, or for the public to check their BGP routing tables for fault diagnosis and further analysis. We can login to these route servers by telnet and run the command "show ip bgp [PREFIX]" to retrieve the BGP routing information of the beacon prefix in these route servers' routing tables. From these routing information, we can get an AS path and other BGP path attributes.

Looking glasses are web pages in which the public can perform `traceroute`, or some of them with "sh ip bgp [PREFIX]" facility. `traceroute` results give us IP routing paths from the looking glasses to the beacon prefix. On the other hand, the " sh ip bgp [PREFIX]" facility on these looking glasses give us the BGP routing information which is the same as route servers.

In order to use `traceroute` to retrieve the routing path from the looking glass to the prefix, we cannot use a prefix only. We need a real destination IP address of that beacon prefix. Thus, we set up a machine with an IP address of that beacon prefix. Then we performed traceroute from those looking glasses to this machine. We got an IP path from the looking glass to the machine we set up in that beacon prefix. However, the analysis of our measurement requires AS paths instead of IP paths. We derive the AS path from these collected traceroute IP paths, and perform IP-to-AS mapping. We map the IP addresses on IP paths to AS numbers that these IP addresses belong to. We retrieve these AS numbers from the existing publicly available Route-Views routing tables.

The IP-to-AS mapping is not perfectly accurate because some of the IP addresses on the actual routing paths belong to Internet exchanges. These Internet exchanges only facilitate a traffic exchange point to other ASes. They usually do not announce their own AS numbers, and these AS numbers are not available in the Route-Views routing tables. As a result, the mapping may not be accurate.

4.1.3 Difficulties

traceroute is still not entirely accurate [35]. traceroute gives us an IP path. In this measurement, we analyze AS paths instead of IP paths. Therefore, we map the IP addresses on the IP traceroute path to AS numbers. We call the mapping resulting paths “IP-to-AS paths”. However, some AS numbers on the IP-to-AS paths cannot be found on the BGP AS paths and the path length cannot represent the AS path length. On the BGP level, some ASes do not prepend their own AS numbers to the AS paths when they send out the BGP updates. These ASes are not announced and cannot be found from the BGP routing table. For example, some of these IP addresses belong to routers of Internet exchanges which only provide facilities for ASes to exchange their traffic. Thus, their AS numbers should not exist on the BGP AS paths. However, we can still have a reconstructed AS path with the reverse traceroute if we identify the Internet exchange IP addresses. Another problem is that the number of ASes, i.e. the path length, on IP-to-AS paths is not accurate. Because the traceroute results contain several routers from the same AS, the IP paths contain multiple IP addresses from the same AS.

Arrangement of route announcement with network operators is needed.

Since it is an operational network, it is dangerous to control the router directly without sufficient experience. Any error may affect the operational network. We must contact the network operator of this network and upstream providers to arrange the route announcement.

4.2 Implementation on RIPE NCC

RIPE NCC has 14 Remote Route Collectors (RRCs) in different locations. We selected three locations which include Sweden, Palo Alto USA, and Italy to implement our active measurement. Each location is one measurement setup. The diversity of locations gives us different results.

An overview of the infrastructure is shown in Figure 4.2.

1. Inside the RIS network (the lower cloud), we use a software, `announcer`, on a Linux PC (`moo.ripe.net`) to announce beacon and control prefixes to the three RRCs, RRC07, RRC10 and RRC14. The `announcer` has set up an iBGP connection to these three RRCs. It announced three beacon prefixes, and control prefixes to three RRCs. In the BGP announcements, we use community values to encode the desired prepending.

2. Each RRC decodes the community values in the route announcement according to a `route-map`. We have set up route-maps on each RRC. A route-map is a table which associates the community value to a prepending length of route announcement. Then the RRC announces the prefix with the specified prepending lengths to the upstream ASes.

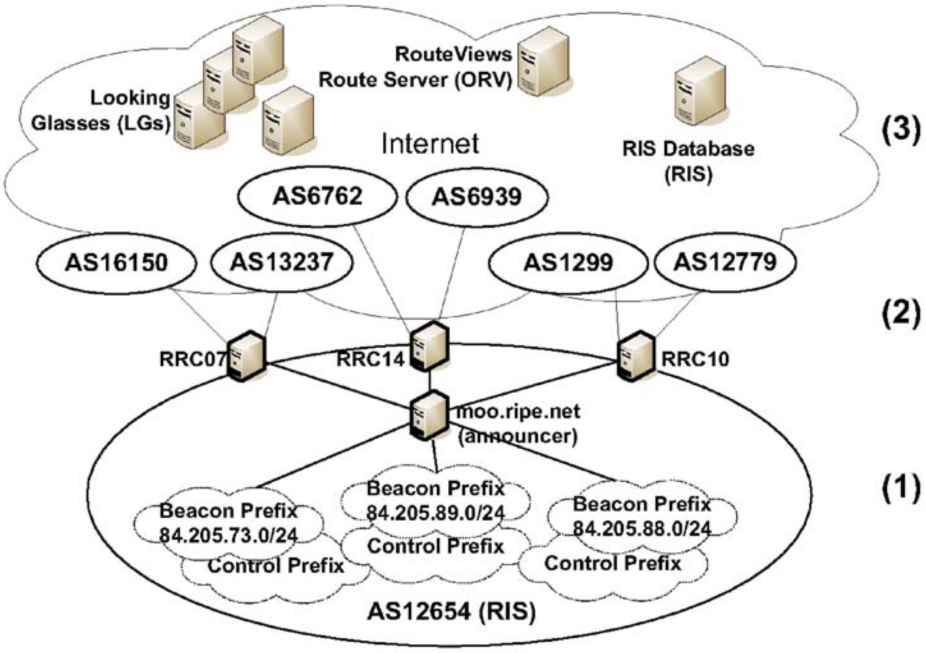


Figure 4.2: An overview of our active measurement infrastructure. (1) Inside the RIPE RIS network (the lower cloud), we use a software, `announcer`, on a Linux PC (`moo.ripe.net`) to announce a beacon prefix with community values encoding the desired prepending to the three RRCs over an iBGP session. (2) Each RRC decodes the community values by mapping them to a `route-map`. The route-map announces the beacon prefix with the specified prepending lengths to the upstream ASes. (3) We observe AS-level route changes from the set of VPs (the upper cloud).

3. We wait for a period for the convergence and observe AS-level route changes from the set of VPs (the upper cloud). The VPs we used in this implementation are looking glasses, Route-Views route servers, and RIS database.

4.2.1 Route Announcement Infrastructure at RIPE RIS

BGP announcements for the beacon prefixes and control prefixes are made by the RRCs of the RIPE NCC RIS project (AS12654) [46]. RRCs are routers or machines equipped with software routers. After performing preliminary evaluations of all 12 RRCs, we setup the route announcement infrastructure at three existing remote route collectors of RIPE RIS [46]: RRC07 (NETNOD, Stockhlm, Sweden) , RRC14 (PAIX, Palo Alto, USA) and RRC10 (MIX, Milan, Italy) to conduct full-scale experiments. The choice was made based on their diversity in geographical location and Internet connectivity: RRC07 is in Stockholm, Sweden, RRC10 is in Milan, Italy, and RRC14 is in Palo Alto, California. Furthermore, each RRC is connected to several upstream ASes. However, in order to reduce the complexity of the measurements and analysis[1], we use only two of the upstreams (i.e., two links) for our measurements. Thus, we are able to perform experiments from different physical locations of the same AS and using different upstream ASes from each physical location.

For each RRC, we prepend only one of the two links, and change the prepending length every two to three hours. We refer to the link that the prepending is applied to as the *prepended link* (PL) and to the other link as the *non-prepended link* (NL). Note that the *maximum prepending length* (MPL) we used for RRC10 is longer than the others because we could still observe noticeable route changes after prepending six times. We have reused some of the beacon and control prefixes when the measurements were done at different times.

We conducted measurement experiments in May 2006. Table 4.1 shows the schedule of our experiments. We sent route announcement and updated prepending length of the beacon prefixes every 2 hours in the RRC07 and

[1] The same experiments can be conducted for more than two links, but the number of possible prepending combinations increases exponentially.

RRC14 measurements. For the measurement on RRC10, we updated the prepending length every 3 hours. Every time before we updated the prepending length, we collected the routing information from VPs.

Unlike Quoitin et al. [41], we do not restart BGP sessions after the convergence of each announcement, as this causes brief disruptions in connectivity and convergence problems due to route flap damping [53] and thus cannot be done in an operational network.

In order to automate the process of route announcement, we use `announcer` and BGP community values [11] to control the prepending length on the announcement on RRCs. The setup is based on: (1) setup on a Linux machine with `announcer` which connects to 3 RRCs and controls the announcement, and (2) setup route-maps to interpret the community value into prepending length of the announcement at RRCs.

Before we implement the measurement on RIPE, we first use a simulator netkit [48] to simulate the Linux machine, 3 RRCs, and the RRCs' upstream providers. We setup the `announcer` and all the route-maps on the RRCs. The simulation was successful and we transferred the whole setup to the network operator to implement the route-maps and `announcer` on the real network.

4.2.1.1 Setup of `announcer` on a Linux Machine

We generate routing announcements by using a Linux PC (`moo.ripe.net`), which maintains iBGP peering sessions with the RRCs using a software `announcer` [17] which is developed by Colitti. This software enables us to change the prepending length of the announcements on different links without modifying the configurations of the RRCs every time. The `announcer` announces the beacon and control prefixes with an empty AS path and a BGP community value. The community value encodes the desired prepending configuration. When the RRCs receive the announcements from `announcer`, they refer to their route-maps about the community value. They prepend their AS numbers on the AS paths according to the route-maps and send out the announcement. By using `announcer`, we do not need to configure each RRC directly for every route announcement. `announcer` allows us to keep configuration overhead to a minimum and avoid configuration errors. Furthermore, it allows us to automate the measurement with simple scripts on the PC to modify the prepending length.

4.2.1.2 Setup of Remote Route Collectors (RRCs) at RIPE RIS

When a RRC receives a route announcement from textttannouncer, it needs to decode the community value on the route announcement. To decode the community value to the prepending length, we configure `route-map` [15] on each RRC. RRCs further announce the route announcements with the respective prepending lengths of the community value to each upstream provider. Moreover, we coordinated with the upstream providers or peers of each RRC about the prepending measurement because some providers' routes filter the prepended routes.

4.2.2 Route Change Collection Infrastructure

We collected possible route changes due to prepending from a set of almost 200 Vantage Points (VPs), including 99 public looking glasses (LGs) from [29], the RouteViews (ORV) route server [4], and the RIS database [46]. From ORV route server and RIS database, we can collect two types of data, BGP routing tables and BGP updates. From LG, we can only collect BGP routing table entries. Note that there may be multiple VPs residing in the same AS: for example, a router may peer with ORV and another host from the same AS may serve as a LG. For the sake of simplicity, we shall refer to any AS that provides its routing information through LG, ORV, or RIS as a VP. BGP routing tables are the routing table snapshots of the routers after they received the BGP updates. On the other hand, BGP updates are produced after a VP updates its routing table and it sends out BGP updates to its peers. Then the peers can notice the route changes and further update their routing table. Because BGP updates have a timestamp to indicate the time when the updates were triggered, we can observe the BGP route selection of the senders and how the updates are propagated.

We have scripts to query these VPs. To obtain routes from the LGs, we developed scripts to parse the HTTP responses. For ORV, we obtain routes from its route server via a telnet connection. ORV has more than 50 peers such that we can have more than 50 views of any prefix. For the RIS, we query the database internally for convenience, but the same information is available via the public web interface. The RIS provides the routes collected by 14 RRCs, which altogether have more than 300 peers, but we can use only 50 of them because not all of RRCs' peers provide full routing tables.

The entire process of collecting the routes from all the VPs is very efficient, taking less than 15 minutes. We query these VPs at least 100 minutes after each route announcement, which is long enough for convergence according to [53]. With our query scripts, we get the routes immediately from LGs and ORV. However, RIS archives the BGP routing table snapshots every 8 hours and the BGP updates every 15 minutes. In order to recover the snapshot at the same collection time of other VPs, we collect the BGP updates and apply them to the 8 hour-snapshot. Then we retrieve the routes of each prefix from the recovered snapshot.

Another source of routing information available to us is the BGP updates made available by the RIS and Route-Views. Because each BGP update has a timestamp on it, we can use the BGP updates to track the evolution of a particular VP's route to the beacon and control prefixes. Furthermore, we can measure the speed with which changes in prepending take effect. For every change in prepending, we can study the number of updates caused and how different ASes responded to it. We present an analysis of this data in §5.2.2.2.4.

4.2.3 Preparation and Difficulties

Because we perform the measurement on a production network, ensuring that the whole setup will not bring down the network is very important. We perform emulations before we implement the measurement. We used `netkit` [48], a network emulator, to test the route-maps on RRCs and `announcer` on `moo.ripe.net`. Netkit provides a variant of Linux kernel and an easy environment to emulate the whole network with only one machine. We first setup the three RRCs, their upstream ISPs routers, and a Linux machine installed with `announcer` on netkit. Then we installed the software routers on these emulated machines and installed the route-maps on them. We test the `announcer` with the configuration scripts. In this case, the route-maps are based on the real architecture. Finally, we implement the whole setup on the real RRCs and `moo.ripe.net`. Organizing the prepending practice with upstream providers or peers is also important. Upstream routers may filter all the routes with repeated AS numbers on it. We notify the upstream ASes before we perform the measurement.

Time of convergence is another issue in the measurement setup because the time between announcement and route collection is dependent on it. Previous work [53] showed that 2 hours were enough for route convergence. Their simulation was only based on new route announcement and withdrawal. However, time of convergence induced by prepending, which is one type of route update, has not been studied. We waited for 100 minutes at the sets of measurements of RRC07 and RRC14. We find that some of the LGs do not have the prefix we announced on their routing table even after 100 minutes. Thus, for RRC10 we announce the preprended route and wait for 3 hours to collect the route changes.

RRC	Upstream ASes		Beacon prefix (Control prefix)	Announcement period
RRC07	AS16150 PORT80	AS13237^{+} LAMBDANET	84.205.73.0/24 (84.205.88.0/24)	8th - 9th May, 2006 (update every 2 hours)
RRC14	AS6762 SEABONE-NET	AS6939^{+} HURRICANE	84.205.89.0/24 (84.205.95.0/24)	8th - 9th May, 2006 (update every 2 hours)
RRC10	AS1299 TELIANET	AS12779^{+} ITGATE	84.205.88.0/24 (84.205.73.0/24)	13th - 15th May, 2006 (update every 3 hours)

Table 4.1: Experiment settings for RRC07, RRC14, and RRC10. We prepend the routes announced to the upstream ASes labeled by $^{+}$. The prepending length is increased from 1, 2 . . . , and up to and including the maximum prepending length.

Chapter 5

Results and Analysis

In this chapter, we first present the overall measurement results obtained from two implementations: one on a local university network which is a dual-homed stub AS, and one that consists of three RRCs of the RIPE RIS. Because these two ASes have completely different topologies and their upstream ASes have different routing policies, the results are different. After that, we highlight a number of specific findings derived from a more in-depth analysis of the routes observed. The analysis helps us to understand not only the effect of prepending, but also predict the results of certain prepending lengths.

Our analysis focuses on the routing dynamics induced by prepending. In our analysis, we classify the upstream ASes on the routes according to their responses to prepending. Some of the upstream ASes respond according to the commercial relationships, while others respond according to their BGP tie-breaking route decisions. Moreover, we discover hidden prepending policies between ASes and some topology issues which are crucial factors to the effect of prepending. From observing BGP updates during prepending, we find that in this topology, more than one link exists between two ASes and these links prolong the time of convergence of prepending.

5.1 Measurement Results

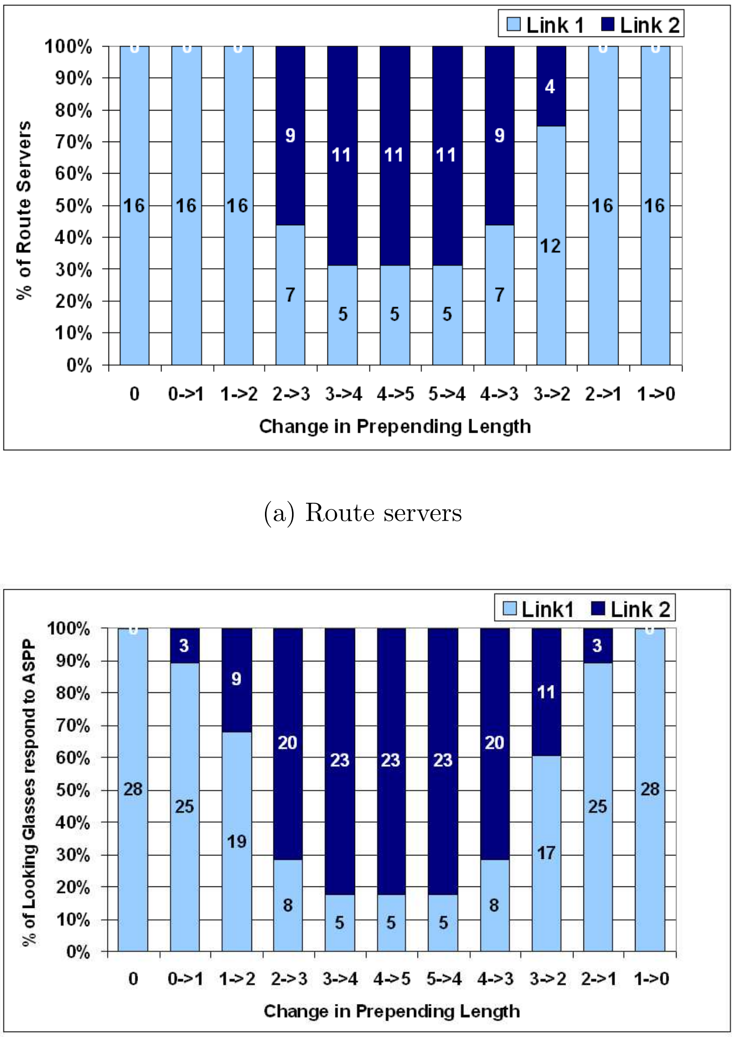
We first present general results for measurements we have done on two different networks on the Internet. We give some statistics of simple cases which happened under prepending. We will present the number of VPs which switched their path because of the prepending in this section.

5.1.1 Measurement Results on the Local University Network

We present the measurement results for the local university network in terms of its link usage. This network is connected to two upstream ASes. We simply use “link” to symbolize the link between an upstream AS and the network. Figure 5.1(a) shows the link usages of the route servers. The x-axis shows the prepending length on link 1 (upstream AS1). $m \rightarrow n$ means the prepending length was changed from m to n to indicate the change of prepending length. The y-axis shows the percentage of route servers that has selected the best route to the beacon prefix via link 1 or link 2. All 16 route servers use link 1 when there is no prepending (prepending length of zero). With a prepending length of one or two, none of them switch to link 2. However, a further increment of prepending beyond two brings an abrupt change in the incoming link for nine of these route servers. Altogether, eleven route servers (68%) respond to the AS path prepending method, and five of them do not. After decreasing the prepending length back to zero, we observe that the distributions of the link usages are not the same for the cases of $1 \rightarrow 2$ and $3 \rightarrow 2$. In other words, the link usage distribution for a given prepending length of two depends on whether the prepending is obtained by forward prepending or backward prepending. Forward prepending means we increase the prepending length and backward prepending means we decrease the prepending length. In general, an unbalancing phenomenon occurs when the link usage distributions are not the same for the cases of $m - 1 \rightarrow m$ and $m + 1 \rightarrow m$, where $m > 1$ is the prepending length.

Figure 5.1(b) shows the link usages by the looking glasses which respond to the prepending. When there was no prepending, there are 28 of them using link 1. There are 14 looking glasses using link 2 and did not respond to the prepending throughout the process. As compared with the route server result, the link changes take place more gradually. Each increment in the prepending length results in link changes for some looking glasses until the prepending length reached four. Altogether, 23 looking glasses (82%) respond to the AS path prepending method. After decreasing the prepending length back to zero, we also observe a slight unbalance between $1 \rightarrow 2$ and $3 \rightarrow 2$.

In our measurement, all the routes are restored to use link 1 after decreasing the prepending length back to zero. Therefore, we do not need to reset the connection to link 2, which is suggested in [23] in order to revert the routes to link 1. However, this may not be true for other ASes.



(a) Route servers

(b) Looking glasses

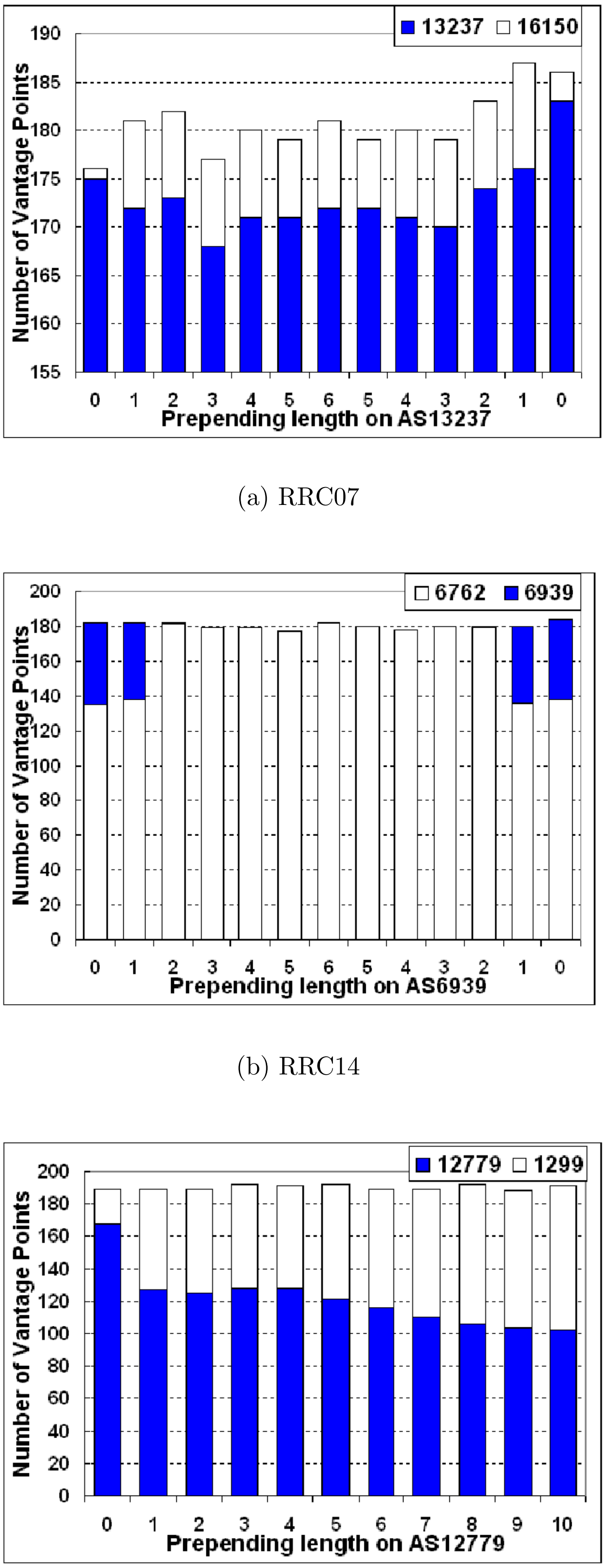
Figure 5.1: The incoming link usages by the route servers and looking glasses under different prepending lengths.

To sum up, the active measurement results in Home AS have shown that the prepending method is quite effective in changing the best routes for the set of sources. This particular set of data also shows that the prepending method can also influence routes when the sources are located farther away from the Home AS. Looking glasses are those located farther away from the Home AS than the route servers. Moreover, the two sets of sources share the commonality that a prepending length of more than four does not cause further route changes.

5.1.2 Measurement Results on RIPE NCC RRCs

In the measurement of RIPE, we chose three RRCs (Remote Route Collectors). These three RRCs, RRC07, RRC14, and RRC10, show three different scenarios of behavior under prepending. On RRC07, we prepend on the busy link, i.e. the link with more paths through it, and observe few changes under prepending. On RRC14, we prepend on the non-busy link and observe a strong effect under prepending. On RRC10, we prepend on the busy link and achieve an intermediate response. Around 60% of ASes changed to the non-busy link after prepending.

Figures 5.2(a), 5.2(b), and 5.2(c) show the link usage under prepending. The x-axis shows the prepending length and the y-axis shows the number of VPs that use the prepended link (PL), which is in blue, and non-prepended link (NL), which is in white, at different prepending lengths. The legends indicate the AS numbers of those PL and NL connected to.



(a) RRC07

(b) RRC14

(c) RRC10

Figure 5.2: The distributions of incoming link usages from VPs at different prepending lengths.

The measurement result for RRC07 in Figure 5.2(a) shows that increasing the prepending length for AS13237 does not have a large effect on routing. However, the total number of VPs continues to change. This is because some of VPs cannot receive the announcement of the beacon prefix and show up "Network not in table" when we query them.

On the other hand, the measurement result for RRC14 in Figure 5.2(b) shows that when the prepending length on AS6939 is increased to two, there is an abrupt change. Nearly all VPs switch to the NL (AS6762). When we further increase the prepending length to three, none of the VPs, including AS6939 itself, uses the PL (AS6939).

We further look at the measurement of RRC10 which we prepend on the busy link, AS12779. Figure 5.2(c) shows that when the prepending length is increased to one, there is already an abrupt change. However, when we further increase the prepending length beyond five, there are still noticeable changes in the link usages. Furthermore, when the prepending length on AS12779 is increased up to ten, more VPs switch to the NL, AS1299. The number of VPs using the links, PL and NL, are almost equal.

Table 5.1 summarizes the effect of prepending in terms of the number of VPs and percentage of ASes that switch from the PL to the NL. We count the number of VPs of the collected routes, and these routes have information of ASes on the AS paths which propagated from the prepended or non-prepended routes. We define the number of ASes here as the number of unique ASes on the routes we collected from VPs, but excluding the prepended AS numbers and AS12654, which is the Home AS. For both of them, we count at time 0,

	RRC07		RRC14		RRC10	
Immediate upstream ASes:	AS13237^{+}	AS16150	AS6939^{+}	AS6762	AS1299	AS12779^{+}
No. of VPs:	184	6	47	139	24	164
No. of ASes:	203	9	55	151	26	180
No. (%) of VPs switched from the PL to the NL:	8 (4%)	-	47 (100%)	-	-	63 (38%)
No. (%) of ASes switched from the PL to the NL:	12 (6%)	-	55 (100%)	-	-	68 (37%)

Table 5.1: The measurement results for RRC07, RRC14, and RRC10 after prepending with their MPLs. "+" indicates prepending on the announcements announced to that upstream AS.

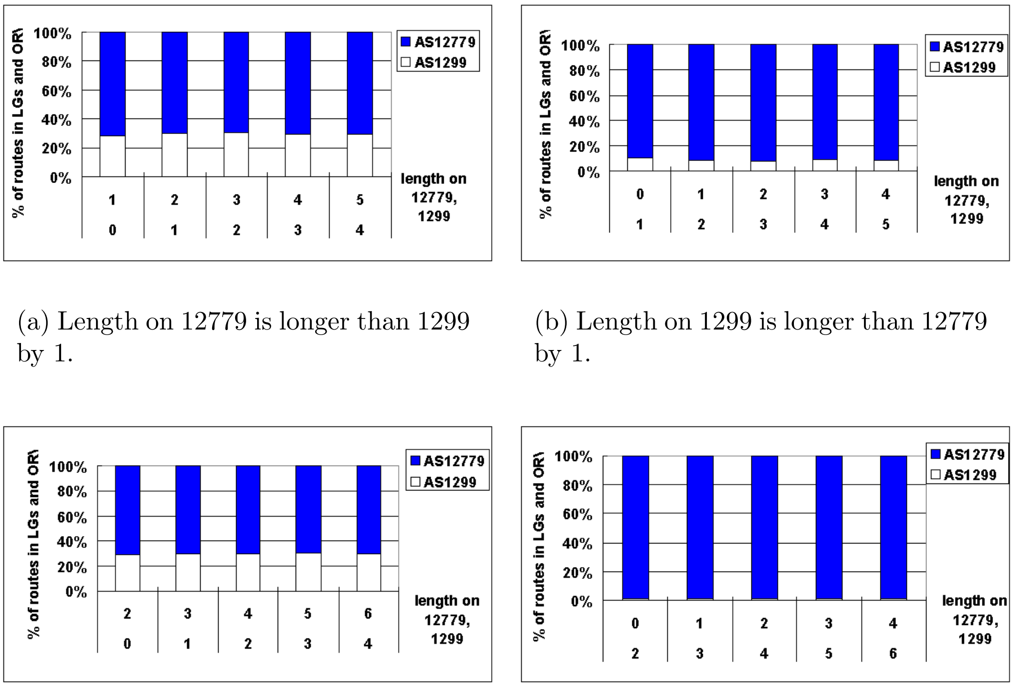
which is no prepending, and afterwards we apply the max prepending length. That is, from time 0 to time 1, we increase the prepending length to 1, and so on. With the AS numbers on the AS paths, we can see the number of ASes using that link without having a VP on those ASes.

Table 5.1 explains the impact of prepending by the number of ASes switched from the PL to the NL. The impact of prepending is very low for RRC07 in spite of prepending on the "busy" link, which is used by the majority of VPs when there is no prepending. In contrast, after the prepending on the "non-busy" link of RRC14, all VPs and ASes switch to the NL, leaving the other link empty. In the case of RRC10, prepending on the busy link has a high impact, switching almost 40% of VPs and ASes to the NL.

5.1.2.1 Prepending on Both Upstream Providers

RRC10 allows us to perform prepending on both upstream providers. We perform this measurement on RRC10. Figure 5.3 shows the results of different prepending lengths between upstream ASes AS12779 and AS1299. The x-axis shows the prepending lengths on AS12779 and AS1299. The y-axis shows the percentage of routes collected from the VPs via AS12779 or AS1299. For each small figure in Figure 5.3, we varied the prepending length for both upstream ASes, but we maintained the same prepending length difference in each figure. We can see for all cases in Figure 5.3, the link usages are the same under the same prepending length difference.

Figure 5.3(a) shows the result of a difference of prepending length 1. The announcement to AS12779 is one prepending length longer than the announcement to AS1299. However, all cases show similar results. We repeat the experiment with prepending on AS1299 one length longer than AS12779 (Figure 5.3(b)), prepending on AS12779 two length longer than AS1299 (Figure 5.3(c)), and prepending on AS1299 two lengths longer than AS12779 (Figure 5.3(d)). All of them show that with the same prepending length difference, they perform the same if we change the prepending length at the same time. It also proves that BGP routers depend highly on the criteria of shorter AS path length to decide which route is the best route.



(a) Length on 12779 is longer than 1299 by 1.

(b) Length on 1299 is longer than 12779 by 1.

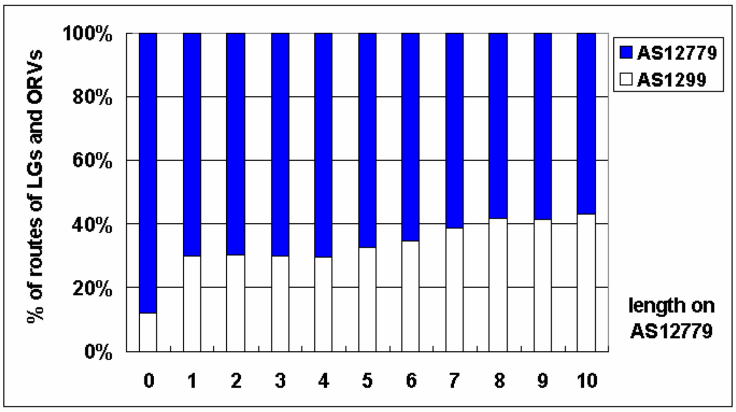
(c) Length on 12779 is longer than 1299 by 2.

(d) Length on 1299 is longer than 12779 by 2.

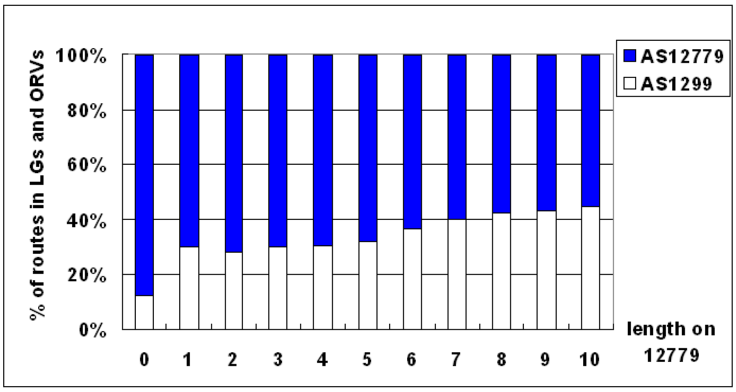
Figure 5.3: The distributions of incoming link usages from VPs at different prepending lengths on both links.

5.1.2.2 Repeatability of the Measurement

We performed the same measurement with the same setup and procedures twice, but at different times. Figure 5.1.2.2 shows the results of two sets of measurements. We increase the prepending length on AS12779 one by one and between the length changes, we have three hours for route convergence. The first time we use prefix 84.205.88.024 (Figure 5.4(a)) and the second time we use prefix 84.205.79.024 (Figure 5.4(b)). Both of them are the beacon prefixes of RIPE NCC RIS. We performed the two measurements on May 12-13, 2006 and May 22-24, 2006, respectively. The figures show that the results are very similar except up to prepending length two, where there are more routes using



(a) May 12-13 on prefix 84.205.88.0



(b) May 22-24 on prefix 84.205.79.0

Figure 5.4: Compare the results of measurements we have done at different times and prefixes.

AS12779 to route to RRC10. Although this simple test does not prove the repeatability of the measurement, it can give us a rough idea on the matter.

5.2 Analysis on the Measurement Results

Besides the statistics, we would like to further understand how and why the prepending induces route changes and affects the routing paths. The measurement result can be used to further understand and perform inbound traffic engineering on the Home AS. Moreover, we classify upstream ASes into several types according to their responses to prepending. With this information, we can understand the network and further perform prepending with information of the responses of upstream ASes. Then we will discuss the results based on BGP mechanics and routing policies. Furthermore, routing convergence is another important issue on inter-domain routing. We will also discuss this issue under prepending which has not been studied yet.

5.2.1 Analysis on the Results of the Local University Network

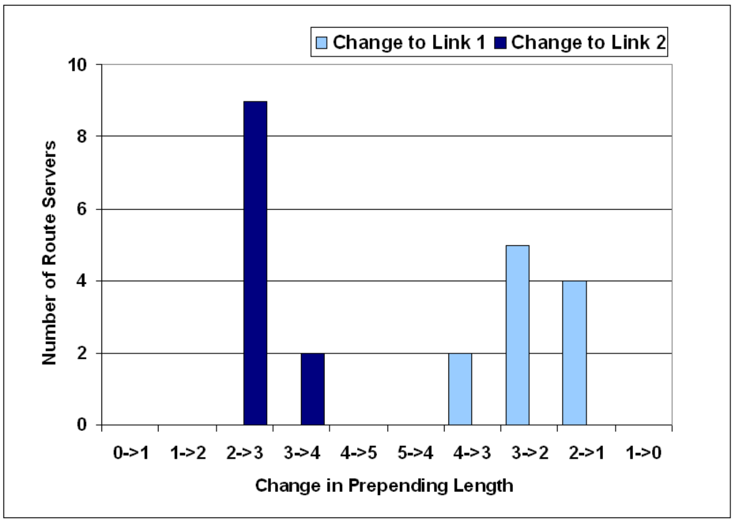
In this section, we will give an analysis on the result of the measurement of the local university network. In this measurement, the number of VPs is relatively smaller than the measurement on RIPE NCC RIS RRCs. Thus the result we discuss here can be specific to this network.

5.2.1.1 How Much Prepending is Needed to Affect the Best Routes?

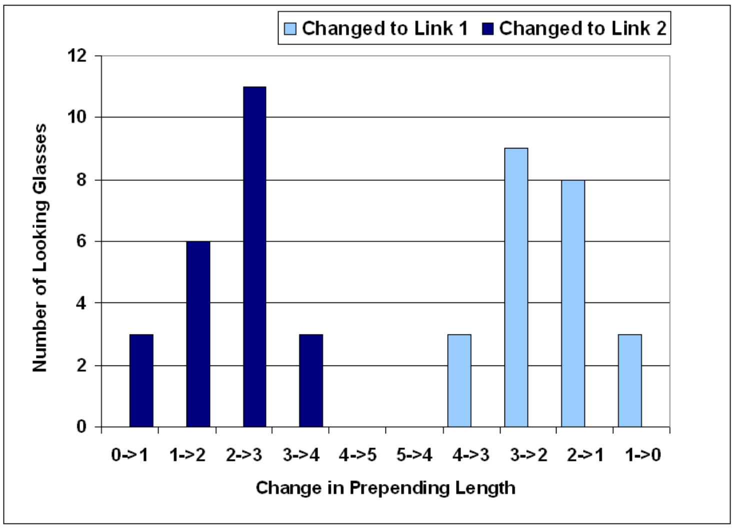
Usually network operators perform prepending on a trial-and-error basis without knowing the exact prepending length they need for their network. With our measurement methodology, we can study the network, find out the prepending length needed, and explain why a certain prepending length is needed for our network.

In Figure 5.5(a) we show the number of route servers that switched their routes at different prepending lengths. As shown, the most significant change takes place when the prepending length is increased to three. However, the changes induced by backward prepending (decreasing the prepending length) are less drastic. The maximum number of route servers involved in the link change in backward prepending is only five, as compared to nine in forward prepending. However, the results of the looking glasses (Figure 5.5(b)) show that the route changes in the forward and backward prepending are more symmetric.

To determine the reason for the abrupt route change on prepending length 2 → 3, we need to focus on the common characteristics of the routes. In Table 5.2, we present more detailed information about the eleven route servers' routes. In the table, we list the common routes from the routes we collected. Route servers shared these AS paths. The second row shows that eleven routes are reaching AS9304 (the Home AS's immediate upstream provider) via AS3491, AS3549, or AS15412. The AS path lengths of these AS paths are at least 3. The route servers are therefore at least three AS hops away from the Home AS. After prepending link 1, all the eleven routes using link 2 share the common AS path: (4637, 3662, 3662, 4528, Home AS). Note that the upstream AS3662 also prepends on this route. Thus, the route servers are at least five AS hops away from the Home AS. In other words, if all the route servers connect to these paths, the prepending length has to be at least two in



(a) The 11 route servers



(b) The 23 looking glasses

Figure 5.5: Distributions of prepending lengths at which route changes take place.

	The AS path	No. route servers
Without prepending (link 1)	(... 3491 9304 Home AS) (... 3549 9304 Home AS) (... 15412 9304 Home AS)	6 3 2
With prepending (link 2)	(... 4637 3662 3662 4528 Home AS)	11

Table 5.2: The routes from the eleven route servers before and after prepending.

order to change the best routes. The lengths of two AS paths differ in two AS hops. As a result, a prepending length of three is sufficient to induce a route change.

From the active measurement results, we have also identified that 45 ASes' import routing decisions are affected by the prepending. That is, they change to use a different route under prepending. Moreover, 32 of them (71%) are three AS hops away from Home AS; only one of them is two AS hops away. The rest are four or five AS hops away. Therefore, most of the route changes take place on the ASes that are three AS hops away from Home AS. Most of the new paths adopted after prepending are longer than the original best paths by two AS hops. Since most of the route changes take place at three AS hops away from Home AS, many prepended routes are not further advertised. As a result, we do not see as many route changes taking place beyond an AS hop distance of three.

Prepending length	No. routes using link 1	No. routes using link 2
0	6 (AS path length = 3)	0
0 → 1	6 (AS path length = 4)	0
1 → 2	6 (AS path length = 5)	0
2 → 3	0	6 (AS path length = 5)
4 → 3	0	6 (AS path length = 5)
3 → 2	3 (AS path length = 5)	3 (AS path length = 5)
2 → 1	6 (AS path length = 4)	0
1 → 0	6 (AS path length = 3)	0

Table 5.3: The routes received by RS_A in the forward and backward prepending.

5.2.1.2 Why is There an Unbalanced Result for the Same Prepending Length?

If the route changes induced by the prepending method are balanced, then the results for the cases $m-1 \rightarrow m$ (forward prepending) and $m+1 \rightarrow m$ (backward prepending) should be the same. However, as shown in Figure 5.1(a) and Figure 5.1(b), this is not the case for both route servers and looking glasses. Therefore, if a stub AS uses AS path prepending to balance the incoming traffic, it is possible to achieve a more balanced link loading using backward prepending instead of forward prepending.

The main reason for the unbalanced phenomenon, as discovered from the active measurement, is due to two non-identical sets of routes VPs receive for cases $m-1 \rightarrow m$ (forward prepending) and $m+1 \rightarrow m$ (backward prepending). That is, the set of routes VPs receive in forward prepending prepending is different compared to the set of routes in backward prepending. Although the prepending length of these cases are m, the set of routes collected are different. With two different sets of routes, the import routing decisions are therefore not necessarily the same.

As an example, we consider one of the route servers, denoted by RS_A. We summarize the routes received by RS_A in Table 5.3 for forward prepending and backward prepending, respectively. Note that the received routes for $1 \rightarrow 2$ are different from that for $3 \rightarrow 2$, although at both times the prepending length is 2. In the case of $1 \rightarrow 2$, RS_A receives six routes, all of which are via link 1 and with an AS path length of five. Therefore, RS_A accepts one of them and continues to use link 1. After the prepending length increased to 3, RS_A switches to routes via link 2. However, in the case of $3 \rightarrow 2$ in backward prepending, RS_A is presented with three routes using link 1 and another three routes using link 2, all of which have the same AS path length of five. Apparently, RS_A decides to continue using a route via link 2. Furthermore, when the prepending length is reduced from $2 \rightarrow 1$, RS_A only receives routes using link 1. As a result, it switches back to link 1.

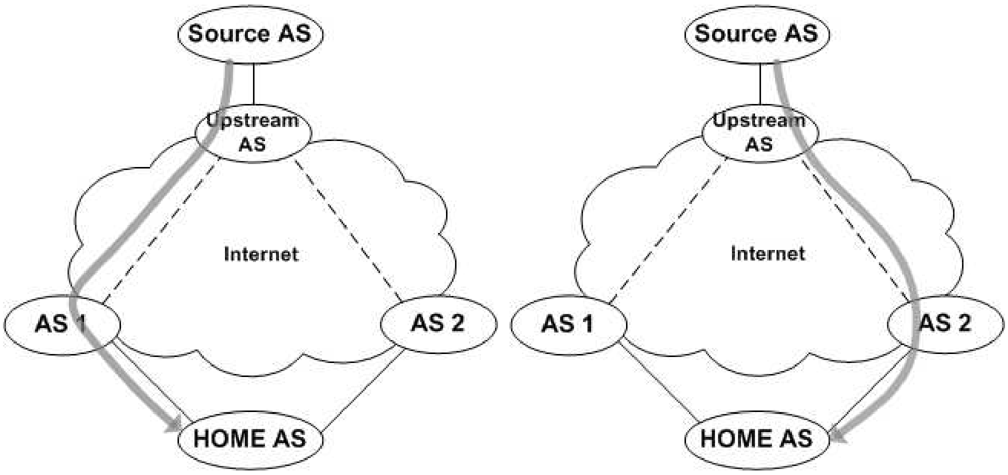
From this example, at a prepending length of 2, both AS path lengths via link 1 and link 2 are the same (i.e., 5). RS_A considers the best route by using tie-breaking rules below shortest AS path length. When the path attributes, such as the AS path length and MED, are the same, the best route will be decided based on other physical constraints of the routes. For example, some of the Cisco routers prefer the route first received (the oldest route) when both routes are from external routers [14]. In Juniper routers, on the other hand, "when a preference tie exists in the same routing table, the physical next-hop of the route with more paths is installed" [27].

5.2.1.3 How are the Best Routes Changed by Prepending?

Based on the active measurement results, we have identified three possible scenarios where a best route can be changed by prepending. With these basic scenarios, we can have a general understanding on how prepending affects the routing paths and the inbound traffic. The first scenario is where the upstream AS of the source AS is solely responsible for the route change (e.g., Figure 5.6). That is, the source AS is not involved in the route change and it passively accepts the route change induced by its upstream AS. Here, the source AS receives the routes to the beacon prefix of the Home AS from its upstream AS and sends traffic to the beacon prefix through the best route it received. One of the examples is that the source AS is singly homed. In this case, the source AS follows the change of its upstream AS.

The second scenario is the opposite of the first one. The source AS is solely responsible for the route change, and the source AS must be a multi-homed AS (e.g., Figure 5.7). That is, the prepending does not affect the source AS's upstream ASes' routes. When the prepended route reaches the source AS, the source AS selects another route as a best route and is solely responsible for making the route change.

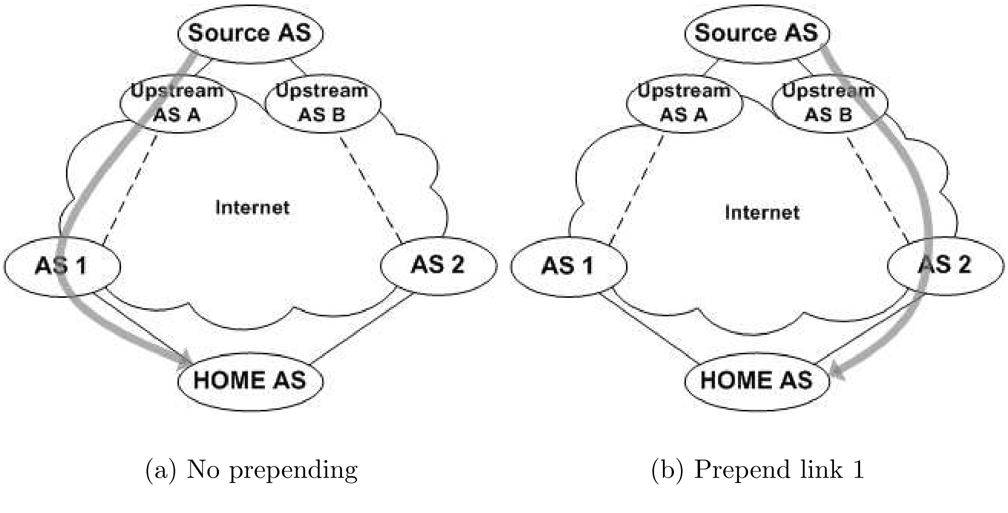
In the third scenario, both the source AS and the upstream ASes are jointly responsible for the route change. In this case, the source AS must be a multi-homed AS. For example, Figure 5.8 shows that before prepending, AS 1 advertises the route via link 1 to AS A and AS B, both of which further announce it to the source AS. The source AS accepts the one announced by AS A because it is the shorter one. After some prepending on link 1, AS B



(a) No prepending

(b) Prepend link 1

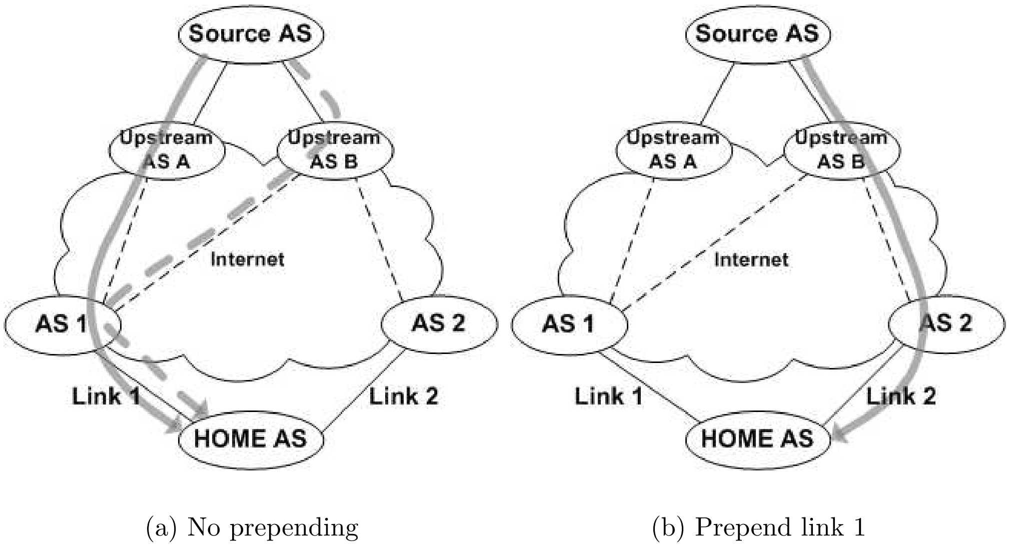
Figure 5.6: Scenario 1: The source AS's upstream ASes are solely responsible for the route change.



(a) No prepending

(b) Prepend link 1

Figure 5.7: Scenario 2: The source AS is solely responsible for the route change.



(a) No prepending

(b) Prepend link 1

Figure 5.8: Scenario 3: The source AS and its upstream ASes are jointly responsible for the route change.

now switches to the route via link 2. Thus, the source AS now receives a route via link 2 from AS *B*. After further prepending on link 1, the route via link 2 becomes the shorter one. As a result, the source AS accepts the route announced by AS *B* instead.

Table 5.4 shows the number of cases according to different scenarios in Figure 5.6, 5.7 and 5.8. The first column shows the number of the first scenario where 26 source ASes follow their upstream ASes to change the routing paths. The second column shows the second and third scenarios together. We can find the third scenario based on VPs on both the source AS and its upstream ASes. If we only have a VP on the source AS, but we do not have a VP on its upstream ASes, we cannot judge whether this route change is solely caused by the source AS alone. In this measurement, 47 cases are source ASes that are solely or partially responsible for the route change. Furthermore, out of 116 ASes, more than half are responsible for the prepending, and around two-

Upstream ASes of the source AS is solely responsible for the route change	Source AS is solely or partly responsible for the route change.	Total number of ASes we found on the data
26	47	116

Table 5.4: The number of cases for different scenarios.

thirds are solely or partially responsible for prepending. These cases infer that most of the ASes consider AS path length in their best route selection. In the route selection algorithm, the ASes first decide the local preference. If the local preferences to different upstream ASes are the same, then they consider AS path length. Thus, if most of the ASes consider AS path length, their local preferences to different upstream ASes are the same.

5.2.2 Analysis on the Results of the Measurement on RIPE

Based on the collected measurement data, we can classify the ASes according to their responses to prepending. We will discuss the classification of direct-responsive ASes and high-impact responsive ASes. Then we will also discuss other issues including commercial relationship, BGP tie-breaking route decisions, hidden prepending policies, and topology. All of these affect the result of prepending. Furthermore, we will study the BGP updates triggered under prepending which will tell us about the time of convergence under prepending.

5.2.2.1 Classifying ASes Based on Their Response to Prepending

We first classify upstream ASes into *responsive* and *non-responsive* ASes. We further note that sometimes in those responsive ASes, we observe that one or more ASes are *responsible* for most route changes. We call them *high-impact responsive* ASes. Identification of these ASes allows greater accuracy in predicting the effect of prepending. We present the classification results and explain the importance of the classification with high-impact ASes.

5.2.2.1.1 Responsive ASes Route changes are mainly caused by *responsive ASes.* Responsive ASes are ASes that switch from the prepended link (PL) to the non-prepended link (NL) after sufficient prepending. Consider the example in Figure 5.9. The arrow shows the direction of route announcements from Home AS. When AS7473 receives route announcements from its next-hop ASes, it can decide which route to use to send traffic to Home AS. Thus, the traffic flows opposite to the arrow. After receiving a prepended route from AS15412, AS7473 changes the next-hop. However, AS7474 does not change its next-hop. Instead, its route change is the result of the route change of its next-hop AS, AS7473. We refer to these two responsive ASes as *direct-responsive ASes* and *indirect-responsive ASes*, respectively.

Not all of the responsive ASes are *responsible* for the changes. We further subdivide responsive ASes into *direct-responsive* ASes (DR-ASes) and *indirect-responsive* ASes (IR-ASes). DR-ASes are responsible for the route change of the IR-ASes. A DR-AS changes its routes such that the next-hop AS on the new route is different. That means it decides to change the route to a NL

The AS path length for reaching Home AS	via link 2				
via link 1	5	6	7	8	9
4	1	**32**	5	1	1
5	0	0	**3**	0	0
6	0	1	0	**2**	1

Table 5.5: The number of direct-responsive ASes for each combination of AS path length for reaching Home AS, via link 1 and link 2.

under prepending. An IR-AS is a responsive AS that is affected by a DR-AS such that it changes its routes. So after prepending, these IR-ASes change their routes to the NL, but have the same next-hop AS on the new route. This is because their next-hop AS changed their routes to NL and these IR-ASes follow their next-hop ASes' decision. Thus, if a DR-AS affects a lot of IR-ASes, it will have a big effect on the prepending.

There are 47 DR-ASes and 26 IR-ASes out of 116 ASes in our measurement. We will now look at their distance from Home AS. Table 5.5 shows the number of DR-ASes and their AS path lengths to Home AS. We have found that the AS path length of the majority of the DR-ASes are four for reaching link 1 and six for reaching link 2. That is, the difference in the AS path lengths for link 1 and link 2 is two. This explains why the greatest route change takes place when increasing the prepending length on link 1 from two to three.

Figures 5.11, 5.12, and 5.13 show the routes announced by the three RRCs. We will use them to explain and discuss some interesting effects of prepending later on.

Consider the three examples in Figure 5.10. The arrows show the direction of traffic, i.e. the preference routes of those VPs to send traffic. The route

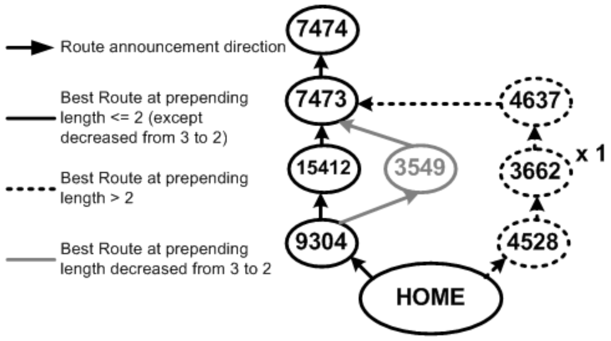
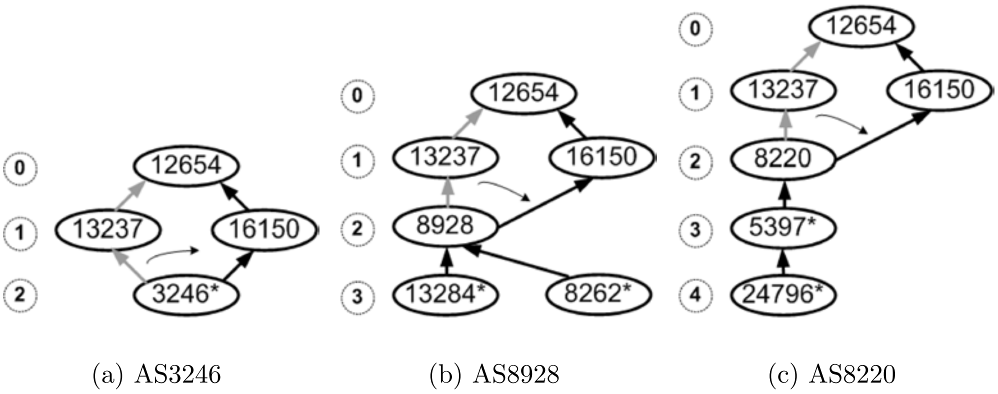

Route announcement direction
Best Route at prepending length <= 2 (except decreased from 3 to 2)
Best Route at prepending length > 2
Best Route at prepending length decreased from 3 to 2
7474
7473
4637
15412
3549
3662
x 1
9304
4528
HOME

Figure 5.9: Example of route change. AS7473 is the direct-responsive AS and AS7474 is the indirect-responsive AS.



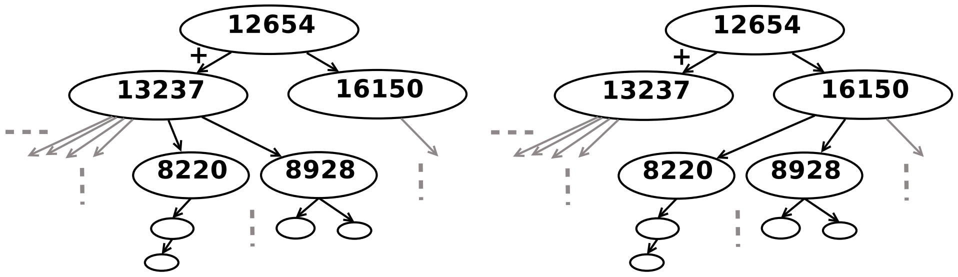
(a) AS3246

(b) AS8928

(c) AS8220

Figure 5.10: Three examples of switching from the PL to the NL. The gray lines correspond to the old routes before prepending; the black lines correspond to the new routes induced by prepending. The AS with $*$ is a VP.

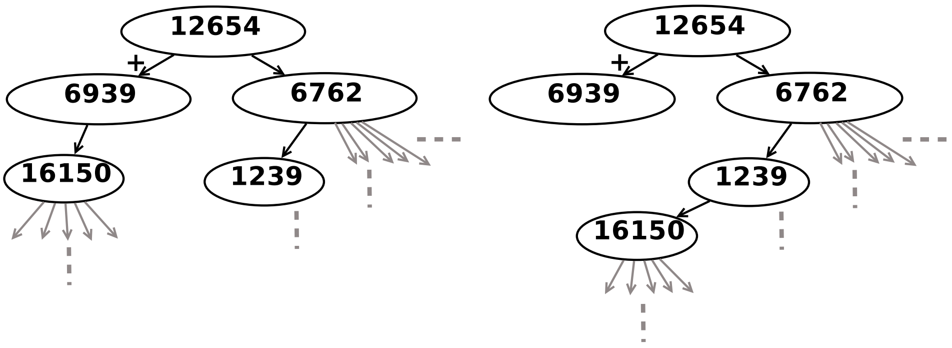
announcements are sent in the opposite directions of the arrows. AS3246 changes its next-hop AS after receiving a prepended route from AS13237 (Figure 5.10(a)). Therefore, we call AS3246 a DR-AS. However, AS13284, AS8262, and AS24796 (Figures 5.10(b) and (c)) do not change their next-hop ASes. Therefore, their switch to the NL (AS16150) is likely the result of a path change of their upstream ASes, AS8928 and AS8220, respectively. We call these ASes IR-ASes. DR-ASes account for 58%, 13%, and 32% of the responsive ASes for RRC07, RRC14, and RRC10, respectively.



(a) Before prepending

(b) After prepending

Figure 5.11: AS paths to the beacon prefix at RRC07.



(a) Before prepending

(b) After prepending

Figure 5.12: AS paths to the beacon prefix at RRC14.

5.2.2.1.2 High-impact Responsive ASes in RIPE Measurement For each DR-AS, there is a set of IR-ASes which are affected by it. Intuitively, we may think of the DR-AS as being at the root of a "subtree" of its IR-ASes. For RRC07, the DR-ASes we observe (e.g., AS8220 and AS8928 in Figure 5.11) map to at most two IR-ASes. For RRC14, AS16150 is a *high-impact AS*, which is responsible for the drastic effect of the prepending method. In fact, AS16150 maps to 46 out of 47 IR-ASes. As we can see in Figure 5.12, the entire subtree of 46 ASes under AS16150 is affected by prepending and changes to the NL. For RRC10, AS12976 is a high-impact AS, as it is responsible for almost 60% of VPs that change to the NL. A more complete topology diagram is shown

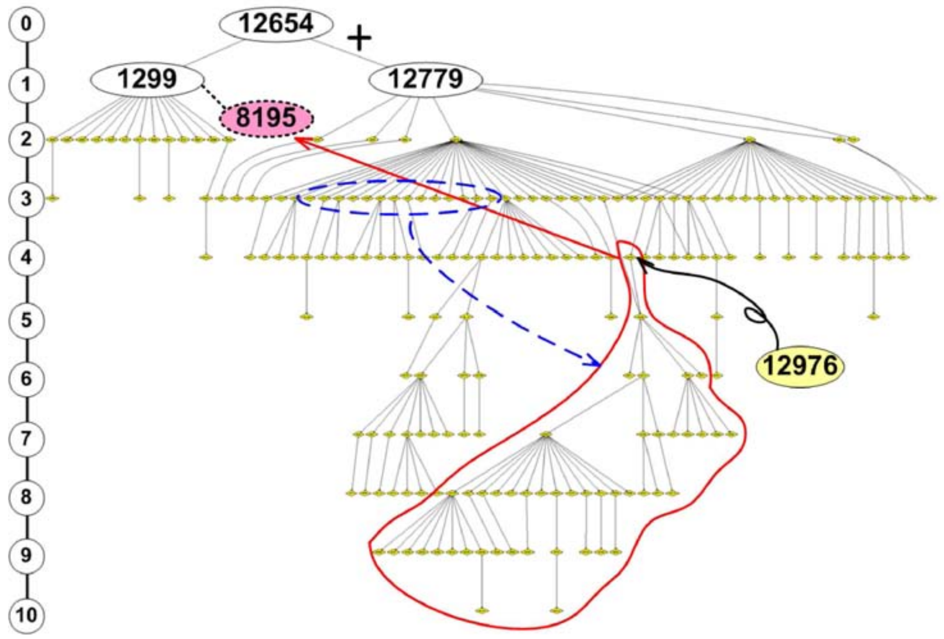


Figure 5.13: The AS-level routing paths to reach the beacon prefixes without prepending for RRC10. Note that AS8195 is not part of the topology before prepending; it appears only after prepending on the link to AS12779.

in Figure 5.13, in which AS12976 and its IR-ASes are enclosed by a solid red line.

5.2.2.1.3 Effect of Prepending to High-impact ASes Past work has observed abrupt changes in terms of inbound traffic distribution at certain prepending lengths (e.g., one in [41] and three in [12]). We have also observed such abrupt changes in terms of the numbers of VPs and ASes for RRC14 and RRC10 at prepending lengths of two and one, respectively. These two “special” prepending lengths are in fact not special at all. They are the respective minimum prepending lengths for AS16150 and AS12976 which are DR-ASes to change their routes. These two ASes are the high-impact ASes in RRC14 and RRC10, respectively. Thus, a high-impact AS dictates the prepending length for which an abrupt route change will occur. Since the topologies of the three RRCs are different, the prepending lengths at which this occurs are different for each RRC.

The complex interaction of routing decisions between high-impact ASes and IR-ASes greatly affects the prepending length distributions. For RRC14, we do not observe route changes beyond a prepending length of three, while for RRC10, route changes are still visible even at a prepending length of ten. The underlying reason is due to a complex interaction of routing decisions between high-impact ASes and IR-ASes. Consider Figure 5.13, at a prepending length of one, AS12976 changes its next-hop AS to AS8195, thus causing route changes in all its IR-ASes and moving its whole subtree of IR-ASes. As we further increase the prepending length, a number of ASes (enclosed by the blue dashed line) move to the AS12976 subtree which already changed to the non-prepended route. In other words, the migration of the AS12976 subtree facilitates other ASes' route changes. As a result of these route changes, the non-prepended routes seen by some of these ASes are introduced after a certain prepending length, thus requiring further prepending to induce route changes. We note that once again, it would be very difficult to obtain this information without the use of active measurements.

Besides identifying DR-ASes and high-impact ASes, it is helpful to understand why some of the upstream ASes are DR-ASes, but others are not. In the following sections we will discuss three factors that contribute the ASes' becoming responsive ASes: commercial relationships between ASes, BGP tie-breaking of the BGP route selection algorithm, and hidden prepending of intermediate ASes. Again, our active measurements allow us to understand how these three issues contribute to the effects of prepending.

5.2.2.2 Other Findings from the RIPE Measurement

With a large scale measurement consisting of VPs, we can analyze the commercial relationships, BGP decision process, hidden prepending policies, and other issues about BGP updates triggered by prepending. We will use the results of the measurement from RIPE NCC RRCs to explain the findings.

5.2.2.2.1 Inferring Commercial Relationships Since DR-ASes respond to prepending, their higher-ranked BGP attribute values, LOCAL-PREF values, must be the same for the routes they used before and after prepending. This means they do not have a higher local preference in these routes. We note that these routes have different next-hop ASes. To explore further, we use the algorithm from [52] to infer the commercial AS relationships of the next-hop ASes of the DR-ASes . For about 40% of DR-ASes, both next-hop ASes are inferred to be providers. That is, those DR-ASes switch from one provider to another provider, and they assign the same LOCAL-PREF value to both providers. The remaining cases are a mixture of sibling-to-peer, sibling-to-sibling, and so on. However, the algorithm fails to identify the relationships for 30% of DR-ASes.

5.2.2.2.2 BGP Tie-breaking Route Decisions Under Prepending A tie-breaking rule is any rule of lower rank than the shortest AS path rule in the route selection algorithm, and is usually not visible outside the AS.

The computer simulation in [41] concludes that reaching the tie-breaking rules in BGP's route decision process is very common. For example, when the local preferences and AS paths of routes from different upstream providers are equal, the tie-breaking rules are used to decide the best path to be used. The tie-breaking rules are not completely decided by the AS itself. For example, the lowest criteria in the route decision algorithm is "Prefer the path that comes from the lowest neighbor address", which is not explicitly set by any AS.

However, our active measurement can help identify tie-breaking route decisions. As an example, consider two VPs in AS3292 and AS15389. AS15389 uses the route via AS3292. The paths used by AS3292 and AS15389 without prepending are (`3292 1239 12779 12654`) and (`15389 3292 1239 12779 12654`), respectively. AS15389 is using a path through AS3292. When prepending length is increased to six (that means the AS path length from AS15389 through AS12779 increases to 11), AS15389 changes its route to (`15389 3292 8342 2118 20483 12976 8195 1299 1299 1299 12654`), which has the same AS path length as the prepended route. Nevertheless, AS3292 continues to use the route via AS12779 until the prepending length is increased to seven. It is possible that AS3292 uses both paths at the same prepending length as a closest egress point [47].

5.2.2.2.3 Discovery of Hidden Prepending Policies and Topology

Our active measurements can also expose hidden prepending policies and hidden ASes. Recall from Figure 5.13 that prepending on AS12779 reveals AS8195. The prepending also uncovers that AS1299 prepends twice the route sent to AS8195, with the goal of discouraging traffic from AS8195. For example, consider the route from AS16150 to AS12654 before prepending: (`16150 8342 2118 20483 12976 1273 1239 12779 12654`). After prepending once on the link to AS12779, AS12976 changes its next-hop AS to AS8195 and the new path is (`16150 8342 2118 20483 12976 8195 1299 1299 1299 12654`). Without prepending, we cannot discover the new path on the topology which has prepending on AS1299. Furthermore, the change in prepending *cancels out* the effect of the prepending inserted by AS1299.

In addition, active measurement can also discover other ASes that are not visible from the VPs without prepending, thus allowing a richer AS-level topology to be discovered. For instance, the prepending announcements for RRC07 increase the number of ASes seen in the topology by 18.7%.

5.2.2.2.4 BGP Updates Triggered Under Prepending To study the effect of prepending on route convergence, we analyze the BGP updates observed during our measurements. The updates were collected from the Route-Views (ORV) archive and the RIPE RIS raw data web site. Whenever an ORV or RIS VP changes its best route to the beacon prefix, it announces the new best route in the BGP updates to the ORV or RIS collectors, which then archive the new BGP updates, into raw data files that we download. Therefore, every announcement we observe for a beacon prefix indicates that a VP has changed its best route to the beacon prefix. Similarly, every withdrawal we observe indicates that a VP has no route to the beacon prefix.

We performed two different experiments to study the BGP updates and route convergence under prepending. First, in order to examine BGP behavior both in the cases of increasing and decreasing prepending length, RRC07 and RRC14 started from a prepending length of zero, incremented by one for each successive announcement up to a maximum prepending length (MPL) of six, and then symmetrically decremented back to zero. Announcements were made every two hours. Second, in order to determine whether convergence time was longer than the two hours used in the other experiments and to investigate behavior with long prepending lengths, RRC10 made announcements every three hours, starting from a prepending length of zero and increasing the prepending length by one with each announcement up to a MPL of 10.

The graphs in Figure 5.14 show the updates collected for the beacon prefixes announced by RRC07, RRC14, and RRC10, respectively. Every + symbol represents an update. The y-axis shows the time in hours from when the first announcement was sent, and the x-axis shows prepending length for the links seen in the update. The position on the x-axis label indicates the prepending length seen in the update. The labels on the x-axis indicate the AS numbers of the non-prepended and prepended link. An update in the shaded column, labeled "W", indicates that the VP sent a withdrawal because it did not have any usable route to the beacon prefix. Updates to the left of the "W" column are announcements using the non-prepended link, and updates to the right of the "W" column are announcements using the prepended link. In order to represent multiple updates observed around the same time for the same prepending length, the updates are slightly offset horizontally from one another in the graphs.

As seen from the graphs, when a new prepending length is announced, we observe BGP updates for both the previous prepending length and the new prepending length. For example, in Figure 5.14(a), at time 2:00 on y-axis, when the prepending length has just been increased to one, we see updates with both prepending length 0 and 1. We believe this is due to the BGP convergence process. Routers that use the prepended route switch to alternate routes with the same AS path length before concluding that the path length has in fact increased by one. These alternate routes have the same AS path, but through different links. Then the new routes arrived and the routers accept the new path with the new prepending length. This is a similar phenomenon observed in [30] for route withdrawals.

We further investigate the updates and find that the new triggered updates with the same prepending length are not identical to those triggered previously. That is, at time 2:00 in Figure 5.14(a), the updates with prepending length 0 are not the same as those at time 0:00 because the MED values of these updates are not the same. MED is one of the tie-breaking processes and its value is used for indicating the preference of link when there is more than one link between two ASes. Table 5.6 shows an example. We collected an update at time 03:10 right after we announced prepending length 1 from AS1239. At 05:10, we announced prepending length 2. But this time, the AS path is the same as the one we collected at 03:10. However, the MED value is 791 which is higher than 1 in the first update (lowest MED value is preferred). Thus, these two updates are not identical because the second one comes from a lower preference of link. It infers that there are at least two physical links between AS1239 and AS1299. One of the links with MED 791 receives the prepended update later than the link with MED 1. Because the MED value is not transitive and can only be passed to the next-hop AS, there are at least

	Time	AS path	MED
1	5/8/2006 3:10	1239 1299 13237 12654 12654	1
2	5/8/2006 5:10	1239 1299 13237 12654 12654	791
3	5/8/2006 5:11	1239 1299 13237 12654 12654 12654	1

Table 5.6: BGP updates of 84.205.73.0/24 collected

two links between AS1239 and AS1299, but not between other ASes on that AS path. After one minute, another new update, which carries the correct AS path length, arrives. The oscillation between two links with different MED value was discussed in [24].

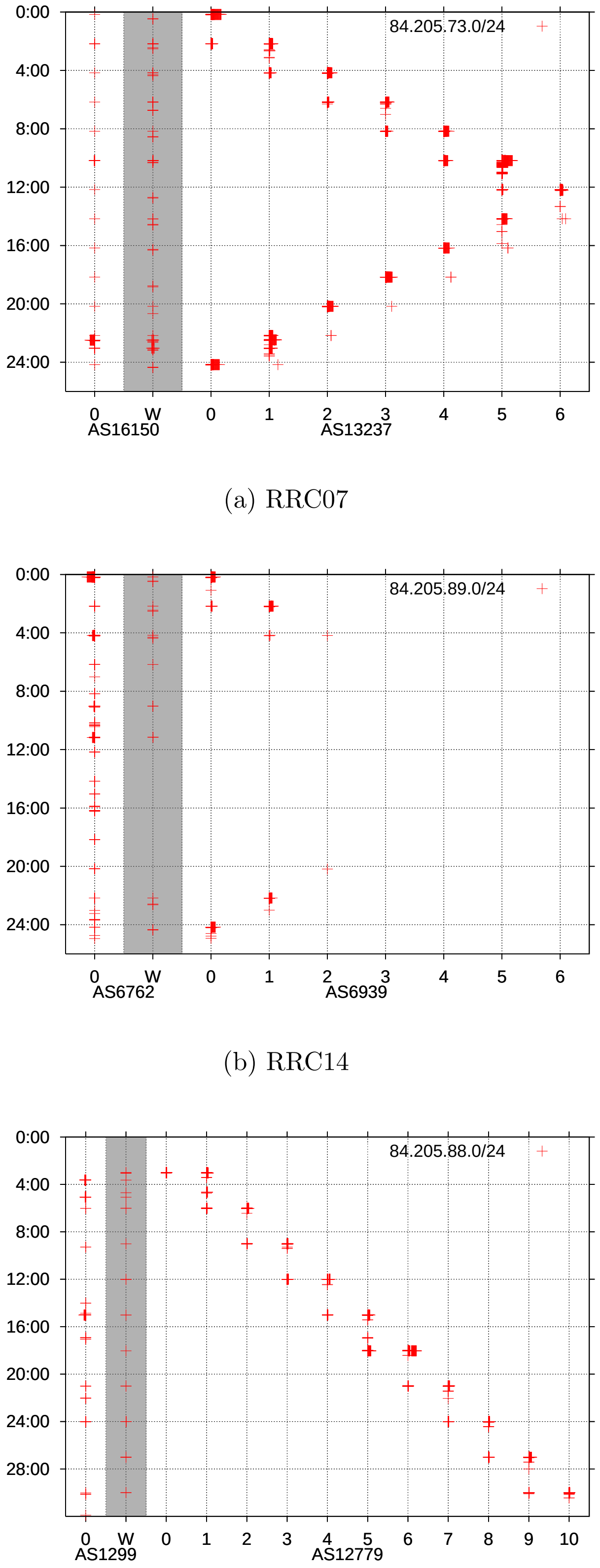
5.3 Comparison Between Two Active Measurement Studies

The two active measurement studies have the following differences:

- Measurement setup:

  We perform the measurement on the local university network by performing prepending on the border router. In this implementation, manual configuration is used to announce the prepended route. However, the measurement performed on RIPE is organized by the tool `announcer` without direct contact with a router. Using `announcer` can reduce errors from mistyping commands and other human errors.

  To collect the route changes in the measurement on the local university network, we retrieve the AS routing path from the IP path, which is



(a) RRC07

(b) RRC14

(c) RRC10

Figure 5.14: BGP updates of the three RRCs against different prepending lengths.

collected from reverse traceroute on the looking glasses. As a result, we cannot observe whether the transit ASes perform prepending because the IP path only tells us the IP addresses of the routers where the path routes through. Thus, the IP path can only tell the actual routing path at that moment, not the BGP level AS path. However, in the measurement on RIPE RIS, we collect route changes from looking glasses and route servers with the BGP utilities on them. With the BGP utilities, we can collect the BGP AS paths and further observe the BGP updates between ASes from the BGP update archives.

- The prepending length to generate the most route changes is different:

  In the measurement on the local university network, the greatest route change occurred when prepending length changed from 2 to 3. However, for RIPE RRC07 and RRC10, the greatest route change occurred at prepending length 1. The greatest route change occurred at prepending length 2 on RRC14. These differences all depend on the topology and AS relationships. Without active measurement study, we cannot observe these route change differences on different networks.

- Unbalanced results are not observed on the RIPE measurement:

  When we performed forward and backward prepending, unbalanced results occurred on the local university network. However, the measurement results on RIPE RRCs do not show this unbalanced phenomenon. The most possible explanation is that in the measurement of the local university network, the upstream ASes had different routing policies or topologies which led to different tie-breaking results in the route selection algorithm. When the AS path lengths of the available routes for the upstream ASes are equal, the routing policy or topology would affect the route selection. We also explained this in Section 5.2.1.2. However, because RIPE RRCs are normally used for collecting measurement and routing information from European networks, preferences of the link usage may not apply to the RRCs. As we can see, active measurement can reveal the policy and observe the response of ASes without knowledge of the upstream ASes.

Chapter 6

Conclusion and Future works

In this dissertation, we have proposed an active measurement methodology for studying routing dynamics induced by AS path prepending. The effect of AS path prepending is difficult to study and understand. Our active measurement methodology allows us to study the impact of AS path prepending on the Internet. Our design minimizes the disruption to normal Internet services and is feasible to deploy in a production network.

The deployments of active measurement have led us to discover a number of hidden processes in the course of propagating prepended routes in the Internet which have not been discussed before. We have deployed the measurement on two stub ASes, the RIPE NCC RIS project and a Hong Kong university network. In the measurement, we observed the link changes due to prepending and studied the upstream ASes according to their behavior to prepending. Furthermore, our methods can help identify tie-breaking route decisions and expose hidden prepending policies and links. Thus, this active measurement is useful for network operators to study their networks before they perform any operations.

We cataloged our measurement data on DatCat.org [19]. We are the first to put this type of measurement result available for other researchers to perform further analysis.

6.1 Future Extension

There are several avenues to extend this work. An important issue to study is the stability of the prepending method. We can study the time of convergence of prepending as a primary metric and this can be analyzed by observing BGP route updates. Once the prepended route is announced, other ASes may or may not change their best routes in their routing tables. They also need to further propagate BGP updates to other ASes. However, if the prepending takes a long time for the routes to converge, it can degrade the performance of the Internet. For this purpose, it would also be useful to increase the number of VPs in order to collect more routes and BGP updates. Thus, we will have a more representative view of the Internet.

On the application side, we can extend the measurement to systematically configure prepending with other traffic engineering methods, such as selective announcement. For example, if we are able to identify a high-impact AS, we know that all routes passing through it will be affected by prepending. Thus, if we perform prepending on one of these high-impact ASes (for example, the transit provider of a stub AS), we may obtain more direct control on traffic switching. Finally, we can investigate the relationship between AS path prepending and end-to-end Internet path performance.

Since Internet resources are limited, optimizing the cost, performance, and available resources is a goal for traffic engineering. This work can be extended with performance measurement and business models. IP network performance of different routing paths can be measured by metrics such as available bandwidth, delay, etc. [25]. One of the extensions is adding PlanetLab nodes to perform the measurement on different routing paths. When the performance of a certain routing path is sub-par, we can perform prepending according to the active measurement results to adjust the inbound routing path. At the same time, we perform optimization with the cost of the routing paths. The optimization can help the current network operators to calculate and adjust the routing paths automatically.

References

[1] African Internet Numbers Registry IP Addresses (AfriNIC). http://www.afrinic.net/.

[2] American Registry for Internet Numbers (ARIN). http://www.arin.net/.

[3] Asia Pacific Network Information Centre (APNIC). http://www.apnic.net/.

[4] Advanced Network Technology Center. University of Oregon Route Views Project. http://www.routeviews.org.

[5] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. RFC3272: Overview and Principles of Internet Traffic Engineering. RFC 3272, May 2002.

[6] G. Battista, M. Patrignani, M. Pizzonia, and M. Rimondini. Towards optimal prepending for incoming traffic engineering. In *Proc. IPS-MoMe*, 2005.

[7] I. Beijnum. *BGP*. O'Reilly, 2002.

[8] A. Broido, E. Nemeth, and k. claffy. Internet expansion, refinement and churn. *European Trans. Telecom.*, January 2002.

[9] T. Bu, L. Gao, and D. Towsley. On characterizing BGP routing table growth. In *Proc. IEEE Global Internet Symp.*, Nov. 2002.

[10] CAIDA. Load Balancing in BGP Environments Using Online Simulation and Dynamic NAT. `http://www.caida.org/workshops/isma/0112/talks/shiv/`.

[11] R. Chandra, P. Traina, and T. Li. BGP communities attribute. RFC 1997, August 1996.

[12] R. Chang and M. Lo. Inbound traffic engineering for multihomed AS's using AS path prepending. *IEEE Network*, pages 18–25, March/April 2005.

[13] CIDR. Report for 27 Jul 07. `http://www.cidr-report.org/as2.0/`.

[14] CISCO. BGP best path selection algorithm. `http://www.cisco.com/warp/public/459/25.shtml`.

[15] CISCO. EIGRP Route Map Support. `http://www.cisco.com/en/US/products/ps6350/product_configuration_guide_chapter09186a0080452963.html`.

[16] CISCO. NetFlow. `http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html`.

[17] L. Colitti. `announcer` software. `http://www.dia.uniroma3.it/~compunet/bgp-probing/announcer`.

[18] J. Cowie, A. Ogielski, B. Premore, E. Smith, and T. Underwood. Impact of the 2003 blackouts on internet communications. http://www.renesys.com/news/2003-11-21/Renesys_BlackoutReport.pdf.

[19] DatCat.org. Internet measurement data catalog. http://datcat.org/.

[20] B. Davie and Y. Rekhter. *MPLS Technology and Applications.* Morgan Kauffmann, 2000.

[21] N. Feamster, J. Borkenhagen, and J. Rexford. Controlling the impact of BGP policy changes on IP traffic. Technical Report 011106-02, AT&T Research, November 2001.

[22] R. Gao, C. Dovrolis, and E. Zegura. Interdomain ingress traffic engineering through optimized AS-path prepending. In *Proc. IFIP Networking Conference*, 2005.

[23] T. Griffin. BGP Wedgies. RFC 4264, November 2005.

[24] T. Griffin and G. T. Wilfong. Analysis of the med oscillation problem in bgp. In *ICNP '02: Proc. of the 10th IEEE International Conference on Network Protocols*, pages 90–99, Washington, DC, USA, 2002. IEEE Computer Society.

[25] IETF. IP performance metrics (ippm). http://www.ietf.org/html.charters/ippm-charter.html.

[26] ISO. Intermediate system to intermediate system routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (iso 8473). ISO/IEC 10589:2002.

[27] Juniper. Examine BGP routes and route selection. `http://www.juniper.net/techpubs/software/nog/nog-baseline/html/verify-bgp9.html`.

[28] S. Kalyanaraman. Load balancing in BGP environments using online simulation and dynamic NAT. Presented at the Internet Statistic and Metrics Analysis Workshops, available from `http://www.caida.org/outreach/isma/0112/talks/shiv/`, December 2001.

[29] T. Kernen. traceroute.org. `http://www.traceroute.org/`.

[30] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet routing convergence. In *Proc. of ACM SIGCOMM 2000*, pages 175–187, September 2001.

[31] Latin American and Caribbean Internet Addresses Registry (LACNIC). `http://www.lacnic.net/en/`.

[32] S. Lo and R. Chang. Active measurement of the AS path prepending method. In *IEEE ICNP '05 (poster paper), available from http://www4.comp.polyu.edu.hk/~csrchang/ICNP05.pdf. Also presented at NANOG 37*, June 2006.

[33] G. Malkin. Traceroute Using an IP Option. RFC 1393, 1993.

[34] Z. Mao, R. Bush, T. Griffin, and M. Roughan. BGP beacons. In *Proc. ACM/USNIX Internet Measurement Conference*, 2003.

[35] Z. Mao, J. Rexford, J. Wang, and R. Katz. Towards an accurate AS-level traceroute tool. In *Proc. ACM SIGCOMM Conference*, 2003.

[36] J. Moy. OSPF Version 2. RFC 2328, 1998.

[37] Number Resource Organization. History of Regional Internet Registries (RIRs). `http://www.nro.net/archive/news/nro.swf`.

[38] K. Patel and S. Hares. Aspath Based Outbound Route Filter for BGP-4. draft-ietf-idr-aspath-orf-09, August 2007.

[39] PlanetLab. home page. `http://www.planet-lab.org/`.

[40] J. Postel. Internet control message protocol (icmp). RFC 792, 1981.

[41] B. Quoitin, C. Pelsser, O. Bonaventure, and S. Uhlig. A performance evaluation of BGP-based traffic engineering. *Intl. J. Network Management*, (15):177–191, 2005.

[42] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, and O. Bonaventure. Interdomain traffic engineering with BGP. *IEEE Commun. Mag.*, 9(3):280–292, May 2003.

[43] Y. Rekhter, T. Li, and S. Hares. A border gateway protocol 4 (BGP-4). RFC 4271, January 2006.

[44] RIPE. RIPE Network Coordination Centre. `http://www.ripe.net/`.

[45] RIPE NCC. bgpdump. `http://www.ris.ripe.net/source/`.

[46] RIPE NCC. Routing information service (RIS) project. `http://www.ripe.net/ris/`.

[47] R. Teixeira, A. Shaikh, T. Griffin, and G. Voelker. Network sensitivity to hot-potato disruptions. In *SIGCOMM '04: Proc. of the 2004 Conference*

on Applications, technologies, architectures, and protocols for computer communications, pages 231–244, New York, NY, USA, 2004. ACM.

[48] The Computer Networks Research Group of the University of Roma Tre and the Linux User Group LUG Roma 3. Netkit. `http://www.netkit.org/`.

[49] F. Wang and L. Gao. On inferring and characterizing internet routing policies. In *Proc. Internet Measurement Conference*, 2003.

[50] F. Wang, Z. Mao, J. Wang, L. Gao, and R. Bush. A measurement study on the impact of routing events on end-to-end Internet path performance. In *Proc. ACM SIGCOMM*, 2006.

[51] H. Wang, R. Chang, D. Chiu, and J. Lui. Characterizing the performance and stability issues of the AS path prepending method: taxonomy, measurement study and analysis. In *Proc. ACM SIGCOMM Asia*, April 2005.

[52] J. Xia and L. Gao. On the evaluation of AS relationship inferences. In *Proc. IEEE GLOBECOM*, 2004.

[53] G. Varghese Z. Mao, R. Govindan and R. Katz. Route flap damping exacerbates Internet routing convergence. In *Proc. ACM SIGCOMM Conference*, 2002.