An Evaluation of BGP-based Traffic Engineering Techniques

L. Swinnen[1], S. Tandel[1], S. Uhlig[2], B. Quoitin[1] and O. Bonaventure[2⋆]

[1] University of Namur, Belgium
Infonet group
[2] Université catholique de Louvain, Belgium
Dept. Computing Science and Engineering

Abstract. We analyze several types of interdomain traffic engineering techniques. First, we briefly describe interdomain routing and the BGP protocol. Then, we summarize the characteristics of interdomain traffic based on measurements with two different ISPs. We evaluate how a typical ISP can select its upstream providers and show that with the BGP decision process many routes are selected non-deterministically. We then evaluate with simulations the performance of BGP-based traffic engineering techniques that are currently used on the Internet and show their limitations.

1 Introduction

This paper studies BGP traffic engineering techniques. The content of the introduction will be written once the other papers for this chapter of the book have been selected.
This paper is organized as follows. In section 2, we briefly describe interdomain routing and the BGP protocol. Then, in section 3, we summarize the characteristics of interdomain traffic. Our main contributions appear in sections 5 and 6 where we analyze interdomain traffic engineering techniques suitable for stub ASes. In section 6, we analyze how a stub AS can control its outgoing traffic and discuss its performance by studying BGP routing tables from various ISPs. Then, in section 6, we present a detailed simulation study of the performance of one technique often used by ISPs to control their incoming traffic.

2 Interdomain Routing

Internet routing is handled by two distinct protocols with different objectives. Inside a single domain, link-state intradomain protocols such as OSPF or IS-IS distribute the entire network topology to all routers and select the shortest path according to a metric chosen by the network administrator. Across interdomain boundaries, the interdomain routing protocol is used to distribute reachability information and to select the best route to each destination according to the policies specified by each domain administrator. For scalability and business reasons, the interdomain routing protocol is only aware of the interconnections between distinct domains, it does not know any information about the content of each domain.

2.1 BGP basics

The current de facto standard interdomain routing protocol is the Border Gateway Protocol (BGP) [RL02,Ste99]. In the BGP terminology, domains are called Autonomous Systems (AS) since these are usually managed by different independent companies. BGP is a *path-vector protocol* that works by sending *route advertisements*. A route advertisement indicates the reachability of a network which is a set of contiguous IP addresses represented by a *network address* and a *network mask* and called a prefix. For instance, `192.168.0.0/24` represents a block of 256 addresses between `192.168.0.0` and `192.168.0.255`. A BGP router will advertise a route to a network because this network belongs to the same AS or because a route advertisement for this network was received from another AS. If a router of `ASx` sends a route advertisement for network N to a router of `ASy`, this implies that `ASx` accepts to forward IP packets with destination N on behalf of `ASy`.
A route advertisement is mainly composed of the address/mask of the network and the `next-hop` which is the IP address of the router that must be used to reach this network. A route advertisement also contains the `AS-path` attribute which contains the list of all the transit AS that must be used to reach the announced network. The `AS-path` has two important functions in BGP. First, it is used to detect routing loops. A BGP router will ignore a received route advertisement with an `AS-path` that already contains its AS number. Second, the length of the `AS-path` can be

⋆ Corresponding author, Email: `bonaventure@info.ucl.ac.be`

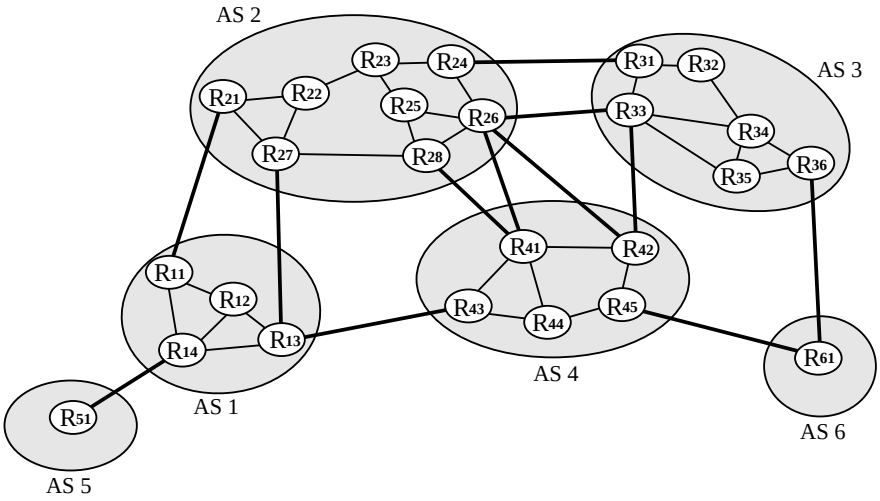

Fig. 1. A simple Internet

considered as the route metric. A route with a shorter `AS-path` will usually be considered better than a route with a longer one.
Besides the `AS-Path`, a route advertisement may also contain several optional attributes such as `local-pref`, `multi-exit-discriminator` (`med`) or `communities` [RL02,Ste99].

2.2 eBGP vs iBGP

There are two variants of BGP [RL02,Ste99]. The eBGP (*external-BGP*) variant is used to announce the reachable prefixes on a link between routers that are part of distinct ASes (e.g. R_{51} and R_{14} in figure 1).
The iBGP (*internal BGP*) variant is used to distribute within an AS the best routes learned from neighboring ASes without injecting interdomain routes into the IGP which is a solution that would not scale in large ASes because of the complexity of computing a shortest path. The iBGP variant is the explanation of how routers of the same AS learn about routes from each other. For this purpose, the basic approach for a BGP router inside an AS is to establish an iBGP session with all the other BGP routers of the same AS. This will result in a full-mesh of iBGP sessions inside the AS. For example, inside `AS1` in figure 1, there will be a full mesh of iBGP sessions involving at least routers R_{11}, R_{13} and R_{14}. These iBGP sessions will be used for example by router R_{14} to announce to the other BGP routers of the AS the route advertisements received from `AS5`.
It is worth noting that a full-mesh of iBGP sessions is not a configuration that will scale well in large AS. There are two different approaches to solve this problem. The first one is to rely on the use of *route-reflectors* which are special BGP routers that will learn/redistribute routes from others BGP routers (sub-sequentially called *route-reflector clients*) within the AS without the need of a full-mesh. The second approach known as *AS confederations* is to divide a single AS in sub-ASes. A slightly modified version of eBGP is used between sub-ASes whereas iBGP is used within each sub-AS. These two approaches are extensively described in [Ste99].

2.3 Route filtering

Inside a single domain, all routers are considered as “equal” and the intradomain routing protocol announces all known paths to all routers. In contrast, in the global Internet, all ASes do not play the same role and an AS will seldom agree to provide a transit service for all its neighbor ASes toward all destinations. Therefore, BGP allows a router to be selective in the route advertisements that it sends to neighbor eBGP routers. To better understand the operation of BGP, it is useful to consider a simplified view of a BGP router as shown in figure 2.
A BGP router processes and generates route advertisements as follows. First, the administrator specifies, for each BGP peer, an input filter (figure 2, left) that is used to select the acceptable advertisements. For example, a BGP router could only select the advertisements with an `AS-Path` containing a set of trusted ASes. Once a route advertisement has been

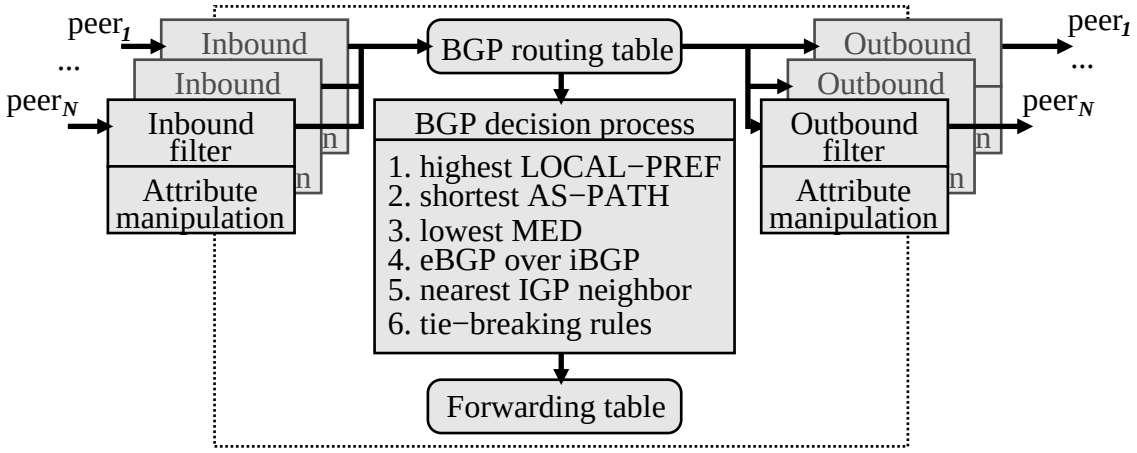

Fig. 2. Simplified operation of a BGP router.

accepted by the input filter, it is placed in the BGP routing table, possibly after having updated some of its attributes. The BGP routing table thus contains all the acceptable routes received from the BGP neighbors.

Second, on the basis of the BGP routing table, the BGP decision process (figure 2, center) will select the best route toward each known prefix. Based on the `next-hop` of this best route and on the intradomain routing table, the router will install a route toward this network inside its forwarding table. This table is then looked up for each received packet and indicates the outgoing interface which must be used to reach the packet's destination.

Third, the BGP router will use its output filters (figure 2, right) to select among the best routes in the BGP routing table the routes that will be advertised to each BGP peer. At most one route will be advertised for each distinct reachable prefix. The BGP router will assemble and send the corresponding route advertisements after a possible update of some of their attributes.

The input and output filters used in combination with the BGP decision process are the key mechanisms that allow a network administrator to support within BGP the business relationships between two ASes. Many types of business relationships can be supported by BGP. Two of the most common relationships are the customer-to-provider and the peer-to-peer relationships [SARK02]. With the *customer-to-provider* relationship, a customer AS pays to utilize a link connected to its provider. This relationship is the origin of most of the interdomain cost of an AS. A stub AS usually tries to maintain at least two of these links for performance and redundancy reasons [SARK02]. In addition, larger ASes typically tries to obtain *peer-to-peer* relationships with other ASes and then share the cost of the link with the other AS. Negotiating the establishment of those *peer-to-peer* relationships is often a complicated process since technical and economical factors, as exposed in [Bar00], need to be taken into account.

To understand how these two relationships are supported by BGP, consider figure 1. If AS5 is AS1's customer, then AS5 will configure its BGP router to announce its routes to AS1. AS1 will accept these routes and announce them to its peer (AS4) and upstream provider (AS2). AS1 will also announce to AS5 all the routes it receives from AS2 and AS4. If AS1 and AS4 have a peer-to-peer relationship on the link between R_{13} and R_{43}, then router R_{13} will only announce on this link the internal routes of AS1 and the routes received from AS1's customer (i.e. AS5). The routes received from AS2 will be filtered and thus not announced on the $R_{13} - R_{43}$ link by router R_{13}. Due to this filtering, AS1 will not carry traffic from AS4 toward AS2.

2.4 Decision process

A BGP router receives from each of its peers one route toward each destination network. The BGP router must then identify the best route among this set of routes by relying on a set of criteria known as the *Decision Process*. Most BGP routers apply a decision process similar in principle to the one shown in figure 2. The set of routes with the same prefix are analyzed by the criteria in the order indicated in figure 2. These criteria act as filters and the N^{th} criterion is only evaluated if more than one route has passed the $N - 1^{th}$ criterion. It should be noted that most BGP implementations allow the network administrator to optionally disable some of the criteria of the BGP decision process.

In most BGP implementations, the set of criteria through which the router goes to select a best route toward a given destination is similar to what follows. First, the router checks that the routes received from its peers have a reachable

`next-hop`, meaning that the IP routing table must contain a route toward this `next-hop`. If more than one route with a reachable next hop exists the router will then use preferences configured by the router administrator. Such preferences may be defined locally to a router with the `weight` parameter or shared over iBGP sessions with the `local-pref` attribute. The router keeps routes with the highest `weight` and then routes with the highest `local-pref`. If after this criterion more than one route remain, the length of the `AS-Path` which acts as the BGP metric is used to compare routes. The length of the `AS-Path` is seen as a measure of the quality of the route and one usually expect that the route with the shortest `AS-Path` is the best.
If at this point the decision process has not yet identified the best route toward the given destination, that means that it has to select one among a set of equal quality routes. The remaining criteria were added for this purpose. The `multi-exit-discriminator` or `med` can be used to compare routes which were received from different routers of the same AS. The route with the lowest `med` is preferred. This criterion is not always enabled because the decision process can be influenced by the remote peers which set the value of the `med`. After the `med`, the decision process prefers routes learned over an eBGP session to routes learned over an iBGP session. The router gives then the preference to routes that can be reached by the closest BGP next hop. If after all these criteria, there is still more than one candidate route, tie-breaking rules are applied. Usual criteria are to keep the oldest route (this minimizes route-flapping) or to prefer the route learned from the router with the lowest ID.

3 Characteristics of Interdomain Traffic

3.1 Source of analyzed data

To obtain a better understanding of the characteristics of interdomain traffic, we have relied on Netflow [Cis99] traces of two different ISPs. Netflow is a traffic monitoring facility supported by Cisco routers. When enabled, the router regularly transmits some information about all layer-4 flows that passed through it to a close-by monitoring station. With Netflow, the monitoring station knows the starting and ending timestamps of all layer-4 flows (TCP connections and UDP flows) as well as the flow volume (in bytes and packets) and the transport protocol and port numbers. Netflow is often used for billing purposes or by ISPs that need to better understand the traffic inside their network. Compared to the traditional packet-level traces that are often analyzed, Netflow has the advantage of being able to monitor multiple links during long periods of time. The main drawback of Netflow is that it does not capture the very short-term variations of the traffic, but this is not a problem in our context of interdomain traffic engineering which tackles medium to long-term traffic variations.
The only characteristics common to both ISPs is that they do not offer transit service. Besides this, they serve very different customers and it can be expected that these customers have different requirements on the network. Due to technical reasons, it was unfortunately impossible to obtain traces from the two studied ISPs covering the same period of time.
The first trace was collected in December 1999 and covers 6 successive days of all the interdomain traffic received by BELNET. BELNET is the ISP that provides connectivity for the research and education institutions located in Belgium. At that time, BELNET was composed of a 34 Mbps star-shaped backbone linking the major universities. Its interdomain connectivity was mainly provided through 34 and 45 Mbps links to the transit service from two commercial ISPs. In addition, BELNET had a 45 Mbps link to the European research network, TEN-155, and was present at the BNIX and AMS-IX interconnection points with a total of 63 peering agreements in operation. Although some universities provided dialup access for their students, the typical BELNET user had a 10 Mbps access link to the BELNET network through their university LAN. During the 6 days period, BELNET received 2.1 terabytes of data. BELNET is representative of research networks and could also be representative of an ISP providing services to high bandwidth users with cable modem or ADSL. We will call BELNET the research ISP in the remainder of this section.
The left part of figure 3 shows the evolution of the total traffic for BELNET during the period of the measurements. While the global evolution of total traffic exhibits a stable daily periodicity, with peak hours located during the day, there are important deviations around the average traffic evolution throughout the day. The mean traffic over the six days period was slightly larger than 32 Mbps, with a one-minute maximum peak at 126 Mbps and a standard deviation of 21 Mbps. The trace begins around 1 AM on a Sunday and finishes six days later around 1 AM also.
The second trace was collected in April 2001 and covers a little less than 5 consecutive days of all the interdomain traffic received by Yucom. Yucom is a commercial ISP that provides Internet access to dialup users through regular modem pools. At that time, the interdomain connectivity of Yucom was mainly provided through high bandwidth links to two transit ISPs. In addition to this transit service, Yucom was also present at the BNIX interconnection point with 15 peering agreements in operation. During the five days of the trace, Yucom received 1.1 terabytes of data. Yucom is representative of an ISP composed of low bandwidth users. We will call Yucom the dialup ISP in the remainder of this section.
The right part of figure 3 presents the total traffic evolution for the dialup ISP during the measurements. The trace starts around 8:30 AM on a Tuesday and finishes almost 5 days later at midnight. The total traffic also exhibits a daily

periodicity with peak hours located during the evening, in accordance with the typical user profile, a dialup user. It had an average total traffic of about 23 Mbps over the measurements, with a one-minute maximum peak at 64 Mbps and a standard deviation of 12 Mbps.

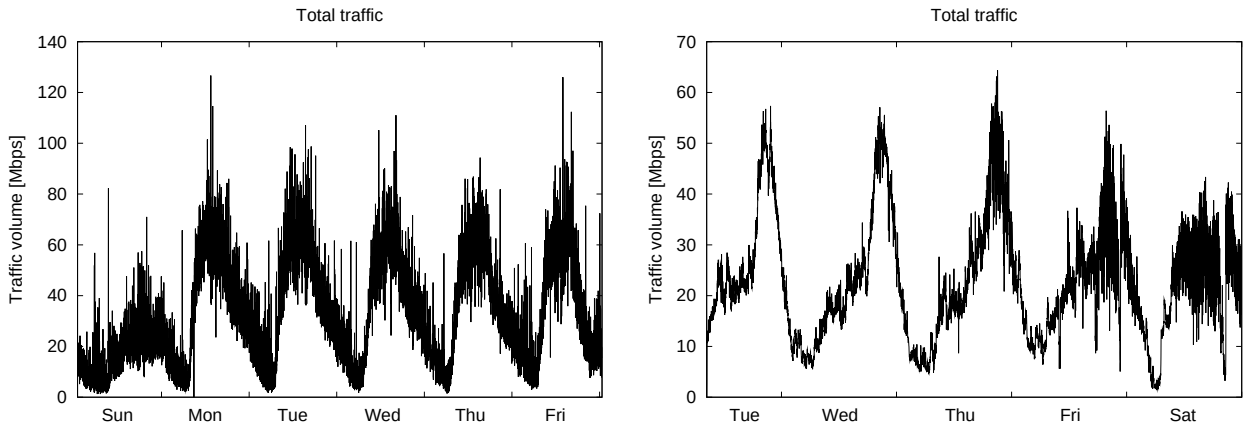

Fig. 3. Total traffic evolution, research ISP (left) and dialup ISP (right).

Before analyzing the collected traffic statistics, it is useful to have a first look at the BGP table of the studied ISPs. In this section, we assume that the BGP table of both ISPs was stable during the period of the measurements and perform all our analysis based on a single BGP table for each ISP. Using a single BGP table for each ISP is an approximation but since we rely on the BGP table of the studied ISPs our analysis is more precise than other studies [FP99,PHS00] that relied on a BGP routing table collected at a different place and time than the packet traces studied in these papers. The routing table of the dialup ISP contained 102345 active prefixes, covering about 26 % of the total IPv4 address space. This coverage of the total IPv4 address space is similar for the research ISP, with about 24 %, but for 68609 prefixes only. Between late 1999 and mid-2001, 30 % more prefixes are necessary to cover a similar percentage of the IPv4 address space. This has already been analyzed elsewhere [Hus01]. Although having different numbers of prefixes in their BGP routing table, the two ISPs cover a similar percentage of the IPv4 address space. This is explained by the average address span per prefix for each ISP, which is about 11000 IP addresses for the dialup ISP and about 15200 addresses for the research ISP. The dialup ISP knew 10560 distinct AS while the research ISP 6298. This difference is mainly due to the large increase in the number of multi-homed sites during the last few years [Hus01]. The average AS path length was 4.2 AS hops for the dialup ISP and 4.5 AS hops for the research ISP.
Figure 4 compares the distribution of the reachable IP addresses for the BGP routing tables of the research ISP and the dialup ISP. The main difference between the two is the more compact distribution for the dialup ISP around a distance of 3 AS hops. The research ISP has its reachable address space more spread over distances of 3 and 4 AS hops. The first 3 AS hops for the dialup ISP provide almost 80 % of the reachable address space while only about 60 % for the research ISP. The difference between the distribution of the reachable IP prefixes seen from the two ISPs is probably due mostly to the 16 months delay between the two traces.

3.2 Topological aggregation of interdomain traffic

To understand the topological variability of interdomain traffic and the possible levels of aggregation, we consider in this section two different types of interdomain flows. Generally, a flow is defined as a set of IP packets that share a common characteristic. For example, a micro-flow is usually defined as the set of IP packets that belong to the same TCP connection, i.e. the IP packets that share the same source address, destination address, IP protocol field, source and destination ports. In this section, we consider two different types of network-layer flows. A *prefix flow* is the set of IP packets whose source addresses belong to a given network prefix as seen from the BGP table of the studied ISP. An *AS flow* is defined as the set of IP packets whose source addresses belong to a given AS as seen from the BGP table of the studied ISP. We do not use explicitly the term “flow” to designate traffic coming from a traffic source, but rather the terms “prefix” and “AS” (or “source AS”) to denote a *prefix flow* and *AS flow* respectively. Note that we use the term *order statistics* throughout this paper to denote the traffic flows ordered by decreasing amount of total traffic sent during the whole measurements.

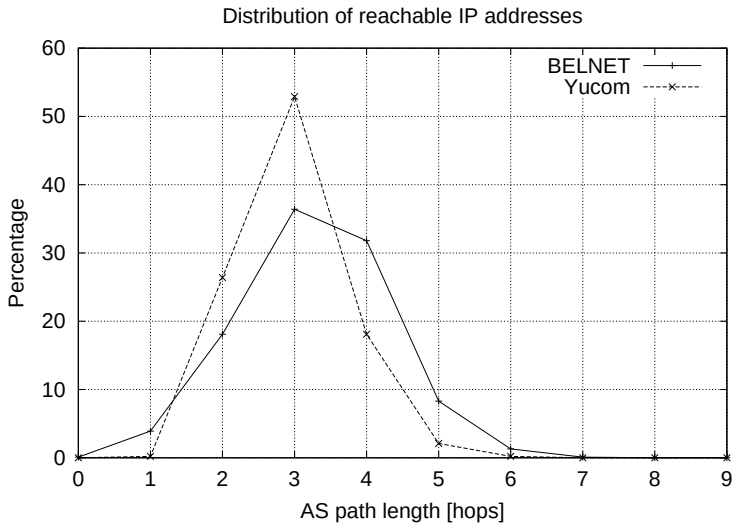

Fig. 4. Distribution of reachable IP addresses.

Let us first study the amount of aggregation provided by the AS and prefix flows. Figure 5 shows the cumulative percentage of traffic for *order statistics* for prefixes and source AS. On this figure, we have thus ordered the prefixes and AS by decreasing order of the total amount of traffic sent by them over the whole measurements, and we have computed their cumulative contribution to the total traffic over the measurements. The x-axis uses a logarithmic scale to better show the low *order statistics*. Both ISPs seem to have a similar distribution for the most important interdomain traffic sources. The top 100 AS (resp. prefixes) capture 72 % of the total traffic (resp. 52 %) for the dialup ISP while a little less than 60 % (resp. a little more than 40 %) for the research ISP. 90 % of the total traffic is captured by 4.7 % of the AS and by 4.1 % of the prefixes for the dialup ISP. The research ISP required 9.8 % of the AS and 4.5 % of the prefixes to capture 90 % of the total traffic. These results are similar to the findings of earlier studies [KN74,CBP93] on the research Internet of the 1970s and the early 1990s. On the other hand, some AS and prefixes contribute to a very small fraction of the total traffic. For the dialup ISP, more than 4000 different AS contributed each to less than 1 megabytes of data during the measurement period and some AS only sent a single packet during this period. For the research ISP, 719 AS sent less than 1 megabytes of data during the six days measurement period.

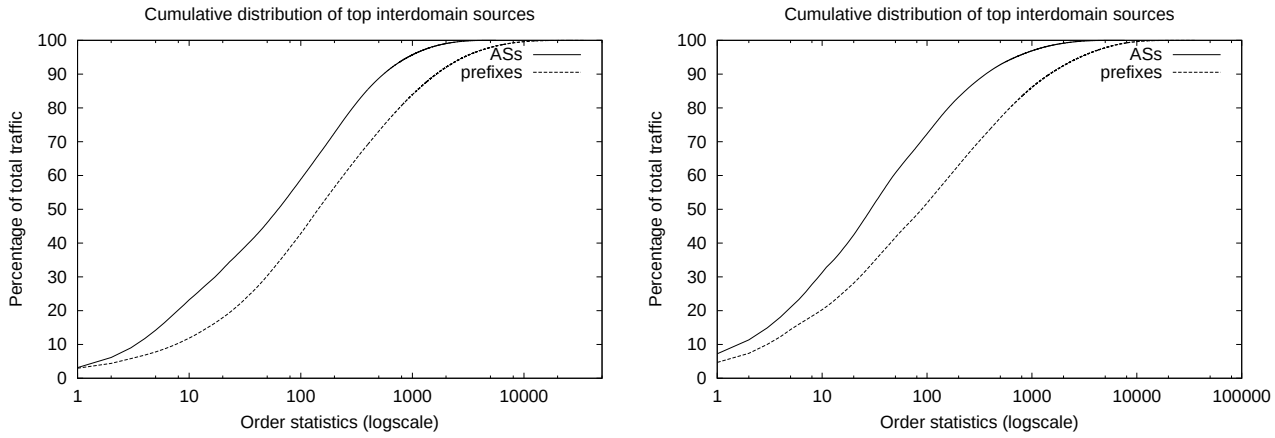

Fig. 5. Cumulative traffic distribution for traffic sources, research ISP (left) and dialup ISP (right).

Another interesting point to mention is that over the measurement period, the research ISP received IP packets from 5606 different AS and 35688 different network prefixes. This corresponds to 89 % of the AS present inside its routing

table. Concerning the dialup ISP, it received IP packets from 7668 different AS and 35693 different network prefixes. This corresponds to 72.6 % of the AS present inside its routing table. These figures show that even relatively small ISPs receive traffic from a very large portion of the Internet during a one week period although some sources only send a few packets.

3.3 Interdomain proximity of the traffic

The amount of aggregation is not the only issue to be considered when studying interdomain traffic characteristics. Another important issue concerns the topological distribution of the traffic. By topological distribution, we mean the distance between the traffic sources and the studied ISP. This distance is important for two reasons. First, usually the performance of an Internet path decreases with the distance between the source and destination AS [McM99]. Second, if the distance between the source and the destination AS is large, it will be difficult for either the source or the destination to apply mechanisms to control the traffic flow in order to perform interdomain traffic engineering [AEWX01].
Figure 6 shows, for each ISP, the percentage of its interdomain traffic that was produced by remote ASes as a function of their distance measured in AS-hops. This figure shows that the studied ISPs only exchange a small fraction of their traffic with their direct peers (AS-hop distance on 1). Most of the packets are exchanged with ASes that are only a few AS hops away. For the BELNET trace, most of the traffic is produced by sources located 3 and 4 AS hops away while YUCOM mainly receives traffic from sources that are 2 and 3 AS hops away.

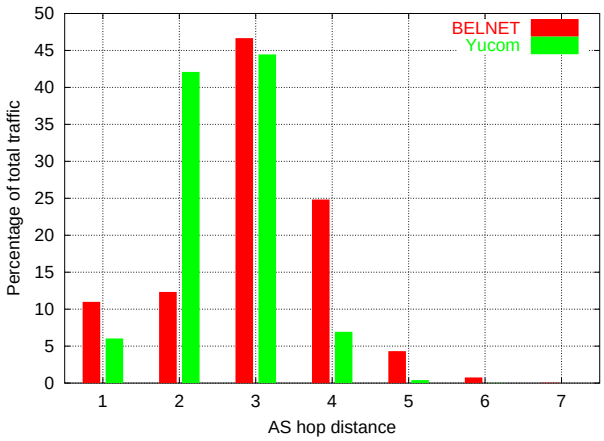

Fig. 6. Per-AS hop distribution of the traffic

3.4 Distribution of the interdomain traffic

The previous section showed the amount of traffic generated by interdomain sources for each AS hop distance. Another concern for interdomain traffic engineering is how many sources send traffic at each AS hop distance. Figure 7 presents the cumulative traffic distribution for the top AS for each AS hop distance. In this figure, an AS is not seen as a traffic source from which a flow originates but also as an intermediate node through which a flow passes. This means that an AS located at an AS hop distance of n is seen as the source of the traffic it generates as well as of all the traffic it forwards when considering the AS_PATH information of the BGP routing table. This means that the traffic seen for all AS at an AS hop distance of n contains the traffic originating from all AS hop distances m with $m \geq n$. Because each AS hop distance does not contribute evenly to the total traffic, we have plotted the cumulative traffic percentage for every AS hop distance with respect to the total traffic seen during the measurements, to show how many AS represent a large fraction of the traffic that crosses the interdomain topology at a given AS hop distance from the local ISP.
The rightmost part of teach curve of figure 7 shows the uneven distribution of the total traffic among the different AS hop distances, equivalent to the information provided by figure 6. The most important AS at 1 AS hop carries 64 % of the total traffic in the case of the dialup ISP while 42 % for the research ISP. This difference is however lessened when

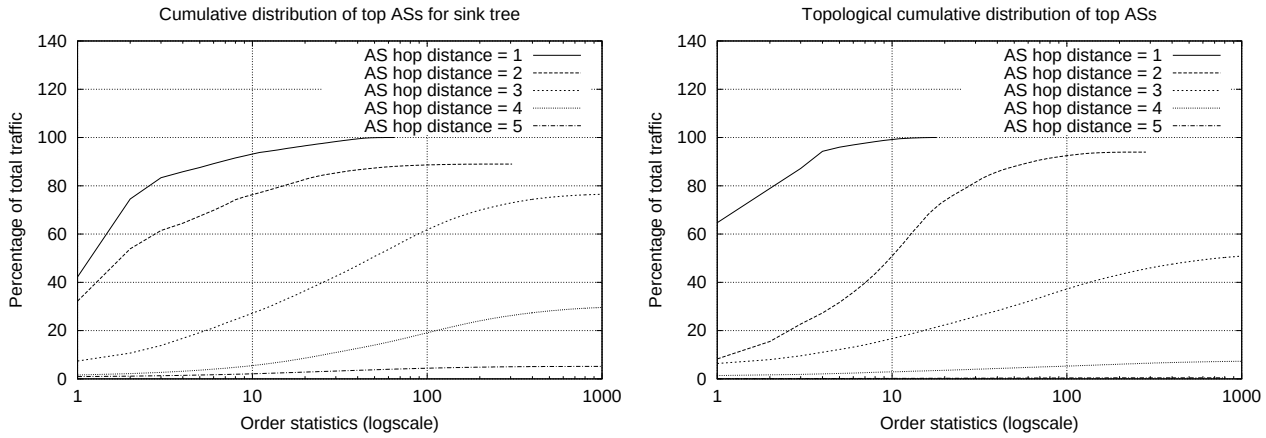

Fig. 7. Cumulative traffic distribution for sink tree, research ISP (left) and dialup ISP (right).

considering the top 3 AS at an AS hop distance of 1, capturing 83 % and 87 % of the total traffic, for the research ISP and the dialup ISP respectively. This shows the predominance of a very small number of BGP peers that provide connectivity for almost all the interdomain traffic of the studied ISPs. At a distance of two AS hops, a few ASes also dominate the traffic with the top 10 carrying more than 77 % of total traffic for the research ISP while 54 % for the dialup ISP. Nevertheless, the traffic produced by AS at a distance of 2 or more AS hops corresponds to 89 % of the total traffic for the research ISP, and 94 % for the dialup ISP. Therefore, a very small fraction of the traffic comes from direct peers themselves. Subsequent distances in terms of AS hops require an increasingly important number of ASes to capture a large fraction of the traffic.
The first AS hop generates 11 % (resp. 6.1 %) of the total traffic, the second AS hop 12.4 % (resp. 42.1 %), the third 46.6 % (resp. 44.4 %), and the fourth 24.9 % (resp. 6.9 %) for the research ISP (resp. for the dialup ISP). The main difference between the two studied ISPs occurs at an AS hop distance of 2. The research ISP has its traffic for the first AS hops that is captured by very few ASes. The dialup ISP on the other hand requires a relatively large number of AS at a distance of 2 AS hops to account for an important fraction of the traffic. This should be compared with the routing table of the dialup ISP (figure 4) where 52 % of the reachable IP addresses are located at 3 AS hops, and 26 % and 18 % at levels 2 and 4. This means that traffic is unevenly distributed between levels 2 and 4, with more traffic coming from level 2 in comparison to its reachable address space relatively to level 4.

4 Interdomain Traffic Engineering

At the interdomain level, ASes have to face various sometimes conflicting issues. On one hand, the traffic is unevenly distributed because BGP seldom takes the right decision on its own and this can cause links to be unevenly loaded and congestion to occur. Moreover, depending on the type of business it handles, an AS will be more concerned by its incoming or outgoing traffic and thus the traffic engineering technique it will use. On the other hand, ASes try to maintain as much connections as they can with other ASes for performance and redundancy reasons. If an AS selects a single provider, then all its interdomain traffic will be sent and received from this provider and the only traffic engineering activity will be to balance the traffic if several physical links are used. However, in practice many ASes prefer, for both performance and economical reasons, to select at least two different upstream providers. Since this connectivity is expensive, another concern of ASes will often be to favor the cheapest links.
Moreover, an AS will want to optimize the way traffic enters or leaves its network, based on its business interests. Content-providers that host a lot of web or streaming servers and usually have several customer-to-provider relationships with transit ASes will try to optimize the way traffic leaves their networks. On the contrary, access-providers that serve small and medium enterprises, dialup or xDSL users typically wish to optimize how Internet traffic enters their networks. And finally, a transit AS will try to balance the traffic on the multiple links it has with its peers.

5 Control of the Outgoing Traffic

To control how the traffic leaves its network an AS must be able to choose which route will be used to reach a particular destination through its peers. Since an AS controls the decision process on its BGP routes, it can easily influence the

selection of the best path. In this section, we first describe two techniques that are frequently used to influence the way the traffic leaves the network. Then we comment the results of an analysis of routing tables collected by Route-Views [oO] which show that an hypothetic stub AS connected to two ISPs often receives two routes toward the same destination. The analysis also shows that the selection of a best route in this set is non-deterministic in many cases (because the lengths of the `AS-Path` are equal).

5.1 BGP-based techniques

A first technique that can be used by an AS to control its outgoing traffic is to rely on the `local-pref` attribute. This optional BGP attribute is only distributed inside an AS. It can be used to rank routes and is the first criterion used in the BGP decision process (figure 2). For example, consider a stub AS with two links toward one upstream provider : a high bandwidth and a low bandwidth link. In this case, the BGP router of this AS could be configured to insert a low `local-pref` to routes learned via the low bandwidth link and a higher value to routes learned via the high bandwidth link. A similar situation can occur for a stub AS connected to a cheap and a more expensive upstream provider.
In practice the manipulation of the `local-pref` attribute can also be based on passive or active measurements. Recently, a few companies have implemented solutions [Bor02] that allow multi-homed stub ASes and content-providers to engineer their outgoing interdomain traffic. These solutions usually measure the load on each interdomain link of the AS and some rely on active measurements to evaluate the performance of interdomain paths. Based on these measurements and some knowledge of the Internet topology (either obtained through a central server or from the BGP router to which they are attached), they attach appropriate values of the `local-pref` attribute to indicate which route should be considered as the best route by the BGP routers. We will evaluate the impact of `local-pref` by using simulations in section 6.2.
A second technique, often used by large transit ISPs, is to rely on the intradomain routing protocol to influence how a packet crosses the transit ISP. As shown in figure 2, the BGP decision process will select the nearest IGP neighbor when comparing several equivalent routes received via iBGP. For example, consider in figure 1 that router R_{27} receives one packet whose destination is R_{45}. The BGP decision process of router R_{27} will compare two routes toward R_{45}, one received via R_{28} and the other received via R_{26}. By selecting router R_{28} as the exit border router for this packet, `AS2` will ensure that this packet will consume as few resources as possible inside its own network. If a transit AS relies on a tuning of the weights of its intradomain routing protocol as described in [FRT02], this tuning will indirectly influence its outgoing traffic.

5.2 Selection of the upstream providers

One of the first interdomain traffic engineering activity of an ISP is to select its upstream providers. The selection of these providers will usually rely on economical criteria, but the BGP routing table of the upstream provider will influence how the ISP will be able to engineer its interdomain traffic.

Measurement study To evaluate the impact of the selection of the upstream provider on interdomain traffic engineering, we have simulated the selection of the upstream provider for a dual-homed stub ISP. For this, we relied on the BGP routing tables collected by route-views [oO]. Route-views is a BGP router that maintains multi-hop BGP peering sessions with 20 different ISPs. We used the BGP routing table recorded on 01 September 2002 at 00:38. We have extracted from the table the routes advertised by each peer of route-views and only consider the peers that advertise their full BGP routing table in this study and used a single BGP peer from each AS. Table 1 shows the list of BGP peers considered and the position of each AS in the Internet hierarchy according to [SARK02].
Based on the BGP routing tables announced by each AS, we have simulated the various possibilities of being dual-homed for a candidate ISP. For this purpose, we performed an experiment with three routers as shown in figure 8. We used three BGP routers and two BGP sessions. Two routers are used to simulate candidate providers and the third router simulates the multi-homed stub AS. Each of the two candidate AS advertises a BGP routing table from one of the providers shown in table 1. The BGP router of the stub AS runs *GNU Zebra 0.92a* [Ish] with a default configuration. This implies that this router selects the best route toward each destination advertised by the candidate ASes based on the normal BGP decision process. We modified *Zebra* to allow us collect statistics on the number of routes selected by each criteria of its decision process.
We used this setup to evaluate all possible pairs of upstream providers based on the route-views data. The first result of this evaluation concerns the size of the routing table of the stub ISP. On average, there were 107789 prefixes inside its routing table with a minimum of 95428 and a maximum of 112842. The upper line of figure 9 shows, for each candidate upstream provider, the average number of routes for the 19 experiments where this provider was considered together with another upstream provider. This figure shows that the average number of routes does not vary significantly.

AS number	Name	Tier level
16150	Port80	3
8121	tch.org	3
1221	Tesltra	2
3130	Randy	5
267	Jared	3
11608	Accretive	3
6539	GT Tel	3
852	Telus	2

AS number	Name	Tier level
2914	Verio	1
3257	TISCALI	2
1239	Sprint	1
7911	Williams	2
3561	C&W USA	1
1668	AOL	4
7018	ATT	1
5511	FT Backbone	1
3549	GLBIX	1
3356	Level3	1
1	Genuity	1
293	ESnet	3

Table 1. Information on the AS tested (ordered by appearance in figure 9)

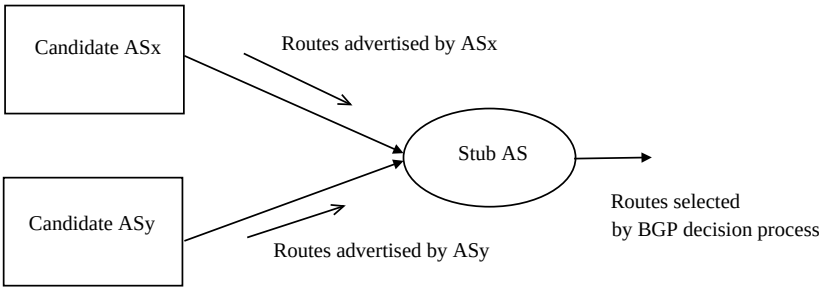

Fig. 8. Experiment to evaluate the quality of the routes from upstream providers

The second element that we considered is the selection of the routes by the BGP decision process of the stub AS. Two cases are possible in our experiment. First, if a route toward a given prefix is only announced by a single candidate AS, this route will automatically be selected. In our experiment, between 87 and 96.5% of the routes received by the stub AS were advertised by both candidate upstream providers. The second line on figure 9 shows, for each candidate upstream provider, the average number of common prefixes between this provider and the other 19 providers. The difference between the routes advertised by different providers can be caused by several factors such as the differences in reachability of each provider and the utilization of prefix-length filters by some AS as discussed in [BBGR01].
The second, and more interesting case to consider is when both candidate providers advertise a route toward each prefix. In this case, the BGP decision process of the stub AS needs to select the best route for each prefix. With the default configuration used by the router of our stub AS, it will first check the `AS-Path` of the received routes. If their `AS-Path` differ, the route with the shortest `AS-Path` will be selected. Otherwise, the tie-breaking rules will be used to select the best route. The bottom line of figure 9 shows, for each candidate upstream provider, the average number of routes from this provider that are selected on the basis of their shorter `AS-Path` by the BGP decision process of our stub AS. For example, concerning AS16150, this figure shows that on average, it advertises 5427 routes with a shorter AS-path than other candidate upstream providers on an average of 102945 routes in common. This is not surprising since this AS is a much smaller than that the tier-1 ISPs found in the right part of table 1.
The second line starting from the bottom in figure 9 shows the average number of routes that were selected by the BGP decision process due to a shorter `AS-Path` in the 19 experiments that involved each candidate AS. For example, concerning AS16150, this line shows that on average, 86104 routes were chosen for their shorter `AS-path` when this provider was confronted with the other providers. The difference between the two bottom lines in figure 9 corresponds to the average number of routes received from the 19 other providers with a shorter `AS-Path` than the considered candidate AS. For AS16150, the other providers advertised on average 80677 routes with a shorter `AS-Path` than this provider. Finally, the difference between the number of common routes and the total average number of shorter routes, shows the average number of routes that were selected in a non-deterministic manner by using the tie-breaking rules of the BGP decision process. For the experiments with AS16150, only 11413 routes were chosen in such a non-deterministic manner.

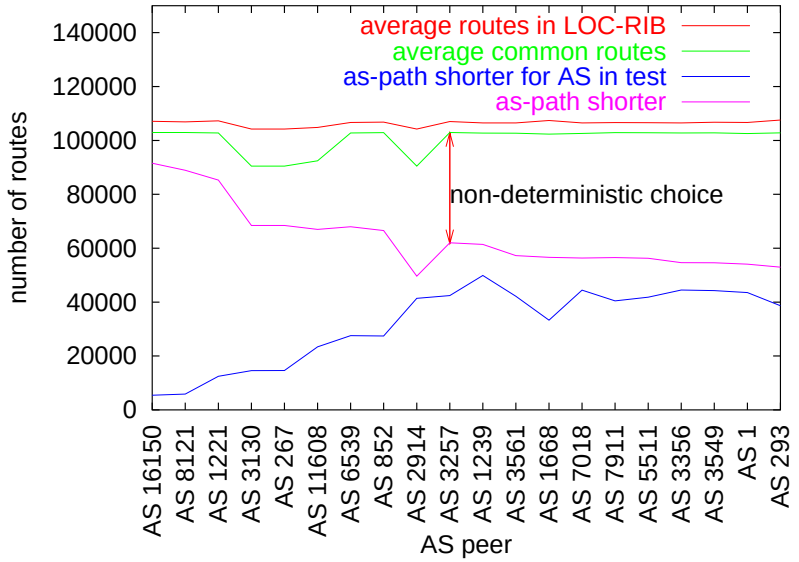

Fig. 9. Quality of the routes announced by an AS — Tests on the 20 peers from Route-Views

If we consider a larger AS in figure 9 such as AS1 (or any of the ASes in the right part of table 1), we find several interesting results. First, on average, AS1 advertises for 43532 prefixes a route with a shorter `AS-Path` than the routes to the same prefixes advertised by any of the other 19 studied ASes. Second, the difference between the bottom line and the line above shows that on average another upstream provider only advertises 10546 shorter routes than AS1. Finally, among all the pairs where AS1 was one of the candidate upstream providers, on average 48473 routes were selected by relying on the tie breaking rules of the BGP decision process. This means that on average 45% of the received routes have the same quality based on their `AS-Path`. Since a stub AS can select any of those routes, this leaves a lot of freedom for interdomain traffic engineering. A closer look at those common prefixes reveals that 50% of the common prefixes with an `AS-Path` length of three AS-hops are chosen in non deterministic manner. Furthermore, 26,5% of the routes chosen by the tie-breaking rules have an `AS-Path` length of 3 or 4. This indicates that the large ASes advertise short routes towards most destinations.

Figure 9 shows us that there are large differences for the ASes in the left and right parts of table 1. This is not surprising since some ASes that peer with route-views like AS3130 are very small and do not serve any customer AS. To better evaluate the large ASes, we performed the same study by considering on the 12 large providers that appear in the right part of table 1.

For these peers that are in majority tier 1, figure 10 shows that the number of common routes is very high varying between 96.9 and 98.1% of the full BGP table except for AS2914 having on average 85% of the routes in common with the 11 other peers. We can also see that the total number of routes chosen on the basis of their shorter `AS-Path` length, is lower than in figure 9. This is because those large ASes advertise shorter routes on average compared to the announcements of smaller ASes. This result also implies that the number of routes chosen by the tie-breaking rules of the decision process is higher than for the previous tests. In fact, figure 10 shows between 56033 and 69735 routes are selected in a non-deterministic manner by the BGP decision process of our stub AS. A closer look at those routes reveals that 80% of them have an `AS-Path` length of 3 to 4 AS-hops. On average, for all considered pairs, almost 62% of the routes are chosen in a non deterministic manner. This result implies that the length of `AS-Path` is not always a sufficient condition to select BGP routes and that ISPs could easily influence their outgoing traffic by defining additional criteria to prefer one provider over the other.

6 Control of the Incoming Traffic

In contrast with the outgoing traffic, it is much more difficult to control the incoming traffic with BGP. Nevertheless access providers can utilize some techniques to influence how the interdomain traffic enters their AS. We first briefly describe these techniques in section 6.1. Then, we present the simulation model that we used to evaluate one of those techniques and discuss the simulation results in section 6.2.

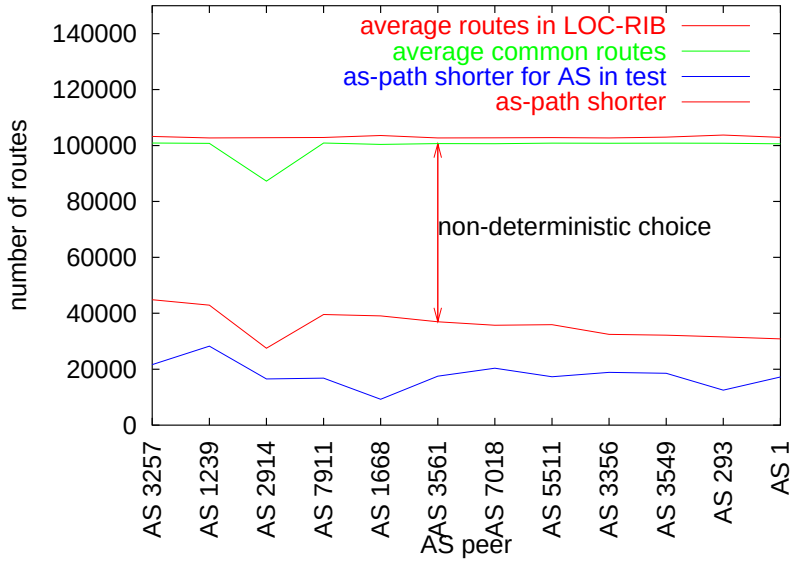

Fig. 10. Quality of the routes announced by an AS — Test on 12 AS

6.1 BGP-based techniques

The first method that can be used to control the traffic that enters an AS is to rely on selective advertisements and announce different route advertisements on different links. This method suffers from an important drawback: if a link fails, the prefixes that were announced only the failed link will not be reachable anymore.

A variant of the selective advertisements is the advertisement of more specific prefixes. This technique relies on the fact that an IP router will always select in its forwarding table the most specific route for each packet (i.e. the matching route with the longest prefix). For example, if a forwarding table contains both a route toward `16.0.0.0/8` and a route toward `16.1.2.0/24`, then a packet whose destination is `16.1.2.200` would be forwarded along the second route. This fact can be used to control the incoming traffic by advertising a large aggregate on all links for fault-tolerance reasons and specific prefixes on some links. The advantage of this solution is that if a link fails, the less specific prefix remains available on the other link. Unfortunately, a widespread utilization of this technique is responsible for a growth of the BGP routing tables. To reduce this growth, many large providers have implemented filters that reject advertisements for too long prefixes [BBGR01].

Another method consists in allowing an AS to indicate a ranking among the various route advertisements that it sends. Since the length of the `AS-Path` appears as the second criteria in the BGP decision process, a possible way to influence the selection of routes by a distant ASes is to artificially increase the length of the `AS-Path` of less preferable routes. This is typically done by inserting several times its own AS number in the `AS-Path`. Based on discussions with network operators, it appears that the amount of `AS-Path` prepending that needs to be used to achieve a given goal can only be found on a trial and error basis.

The last method to allow an AS to control its incoming traffic is to rely on the `multi-exit-discriminator` (MED) attribute. This optional attribute can only be used by an AS multi-connected to another AS to influence the link that should be used by the remote AS to send packets toward a specific destination. It should however be noted that the utilization of the `MED` attribute is usually subject to a negotiation between the two peering ASes and some ASes do not accept to take the `MED` attribute into account in their decision process. Furthermore, the utilization of this attribute may cause persistent oscillations [GW02].

6.2 Evaluation of `AS-Path` prepending

As described in the previous section, `AS-Path` prepending can be used by an ISP to control the flow of its incoming traffic by announcing on some links routes with an artificially long `AS-Path`. Although this technique is used today in the Internet ([BNC02] reports that `AS-Path` prepending affected 6.5 % of the BGP routes in November 2001), there has not been any analysis of its performance to the best of our knowledge.

Simulation Model The first element of our simulation model is our simulation environment : Javasim [Tya02]. Javasim is a scalable event-driven simulator developed by Hung-Ying Tyan and many others at Ohio-State University. Javasim is written in Java for portability reasons and contains realistic models of various Internet protocols. Although Javasim supports several routing protocols, it did not contain any BGP model. Instead of developing a BGP model from scratch, we choose to port[3] and enhance the BGP implementation developed by B. J. Presmore [Pre01] for SSFNet [CNO99]. This model has been extensively validated and tested and has already been used for several simulation studies [GP01,MGVK02]. We have enhanced it to better support the routing policies that are often used by ISPs as shown earlier.

The second element of our simulation model is the network itself. Since our goal is to evaluate the performance of interdomain traffic engineering, we need a realistic model of the Internet. To evaluate `AS-Path` prepending, we choose to build each AS as composed of a single router that advertises a single IP prefix. This router runs the BGP protocol and maintains BGP sessions with routers in neighboring ASes. The second element that we needed to specify is the topology of the interdomain links.

An interdomain topology could be obtained from a snapshot of the current Internet, such as the one analyzed in [SARK02]. However, a drawback of this approach is that then it is difficult to perform simulations with various topologies to evaluate the impact of the topology on the results. A second method is to rely on a topology built by topology generators. Various topology generators have been proposed and evaluated in the last few years (see [FFF99,TGJ02] and the references therein). It is admitted that two classes of generators can be used: structural and degree-based generators. Structural generators attempt to reproduce the real Internet hierarchy (i.e. tiers, transit ASes and stubs) while degree-based generators approximate a specific property of the real topology, the node degree distribution. It has been shown in [FFF99] that the Internet hierarchy can be better approximated with topologies produced by degree-based generators. Moreover, [TGJ02] indicates that degree-based generators are also better suited to approximate the structure of the Internet. Indeed, such generators implicitly create hierarchies closely related to the current Internet hierarchy.

We relied on a degree-based topology generator to produce the various Internet topologies used for the simulations. Our topologies have been generated with Brite [MAMB01] which is a highly configurable generator. One of its interesting features is the ability to produce topologies with ASes only, intended to simulate the interdomain level. Brite is able to rely on various mathematical models to generate a topology. We have chosen the Barabasi-Albert model [BA99] because it is degree-based. This model builds the topology sequentially by adding one AS at a time while relying on two simple principles[AB02]:

- *Growth*: each node that must be added to the topology is connected to m existing nodes (where m is a parameter of the generator).
- *Preferential Attachment*: when a new interdomain link is created, it connects the AS being added to an existing AS. This AS is selected with a probability which depends on the number of links already attached to each AS. This means that an AS with a lot of interdomain links will be attached to other ASes with a high probability.

A consequence of these two principles is that the ASes which are generated first (i.e. those with a low identifier) have a greater connectivity than the ASes generated last (i.e. those with a large identifier).

In our simulations we use an interdomain topology with two types of ASes. The *core* of the network is composed of a few hundred transit ASes. This core is generated by using Brite. Note that we do not consider hierarchy in the *core*. For this reason, all *core* ASes all advertise their full routing table to their neighboring ASes and no routing policies have been defined for the *core* ASes.

In addition to the *core*, our topologies also contain a few hundred stub ASes. Those stub ASes are added to the topology generated by Brite by following a *preferential attachment* principle (the probability for an AS in the *core* to be connected to a stub is function of its current connectivity). Each stub has exactly two connections to two different transit ASes in the *core*. These connections represent *customer-to-provider* links where the stub is the customer and the ASes in the *core* are the providers. We configured BGP policies on the stub ASes to ensure that those ASes do not provide any transit. In the following, we call `lowID` provider (resp. `highID` provider), the provider of the considered stub AS with the lowest (resp. highest) AS number and `lowID` link (resp. `highID` link), the link that leads to this provider.

We have introduced the stub ASes in our network topologies for two reasons. First, they represented 85.6% of the number of ASes on the Internet in October 2002 $checkdate$. Second, they will serve as measurement points to evaluate the impact of `AS-Path` prepending on the routes selected by the BGP decision process of each simulated router.

We have performed simulations with several topologies. Due to space limitations, we restrict our analysis in this paper to two representative topologies. The first topology is composed of a lightly connected *core* with 200 ASes. This topology was produced by Brite with the value of the m parameter set to 2. Figure 11 (left) shows the topology of the *core* where each square represents an AS and each link a peering link. 400 dual-homed stub ASes were attached to the *core* ASes with preferential attachment. In total, when considering both the stub and the transit ASes, this topology contains 1594 interdomain links.

[3] Our modifications to Javasim will be available soon from `http://www.javasim.org`.

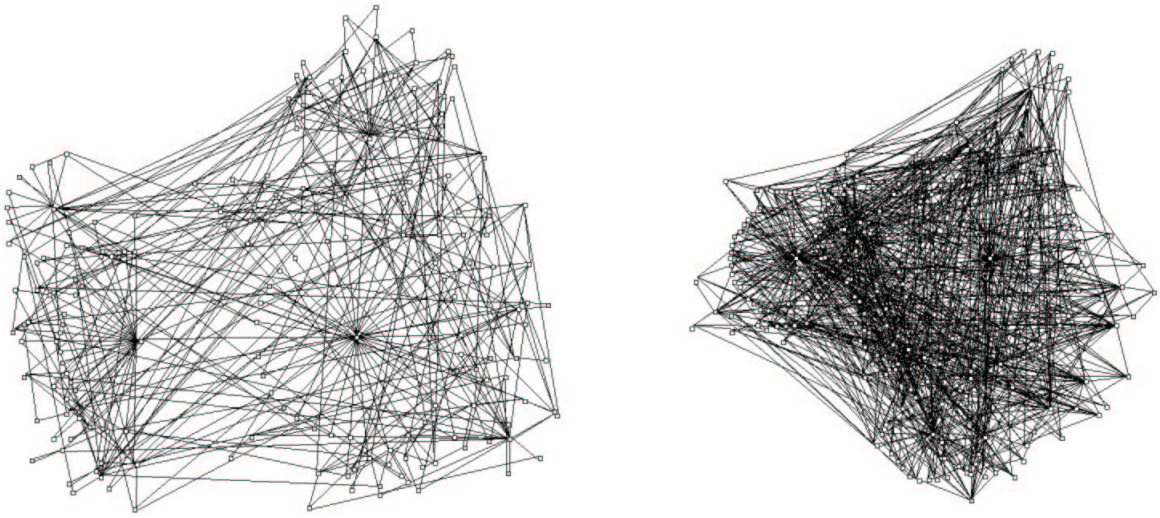

Fig. 11. Interdomain topologies : lightly (left) and dense cores (right)

The second topology is composed of a dense *core* with 200 ASes. This *core* was produced by Brite with the value of the m parameter set to 4 to model a *core* composed of ASes with a higher connectivity. This core is shown in the right part of figure 11. As can be seen, the core network is dense. We have connected 200 dual-homed stub ASes with preferential attachment to this dense core. In total, this second topology contains 1980 interdomain links.

Due to the memory constraints we were not able to perform simulations with more than about 2500 BGP peering sessions. This is one order of magnitude less than the number of interdomain relations reported in [SARK02], but more than one order of magnitude more links than existing traffic engineering studies.

Another element that should be considered in such a model is the amount of traffic sent by each AS towards each remote AS. In section 3, we have shown that a small number of ASes were responsible for a large fraction of the traffic received by the two studied ISPs. However, those measurement are not sufficient to allow us to determine the behaviour of hundred different ASes and how their traffic would be distributed among the Internet. Developping such a model is outside the scope of this paper and for the simulations described below, we will consider the interdomain paths between all AS pairs without considering the amount of traffic exchanged.

Simulations without `AS-Path` prepending To evaluate the impact of this BGP traffic engineering technique, we use the stub ASes as measurement points. After a sufficient time to allow the BGP routes to converge, each router sends a special IP packet with the `record route` option toward each remote stub AS. The stub ASes collect the received IP packets and by analyzing the `record route` option of each received packets, we can determine all the interdomain paths followed by IP packets. The analysis of all these interdomain paths allows us to study the impact of BGP traffic engineering techniques.

For our first simulation, we configured the stub ASes to send their BGP announcements without any `AS-Path` prepending. Figure 12 shows, for each stub AS, the percentage of the interdomain paths that end on this stub AS and are received via its `lowId` provider. To plot this figure, we have ordered the stub ASes that appear on the x-axis in decreasing percentage of the interdomain paths received via their `lowID` provider. This ordering, determined for each topology, is used for all simulation results described in the remainder of this paper.

Several points need to be mentioned concerning figure 12. First, the distribution of the interdomain paths is not uniform. For both core networks, some stub ASes receive almost all their interdomain packets via one of their providers. The stub ASes that receive almost all their interdomain packets via their `lowID` provider are usually attached to a dense `lowID` provider and a `highID` provider with a very weak connectivity. For the stub ASes that receive almost all their interdomain packets via their `highID` provider, the reason is that this provider is closer to the core ASes with the higher connectivity than their `lowID` provider. Note that with the dense core this situation occurs less often.

A second point to be mentioned is that for the lightly connected core, about 66% of the stub ASes receive more than 60% of their interdomain packets via their `lowID` provider. This is due to the fact that, thanks to its connectivity, the `lowID` provider is, on average, closer to most destinations than the `highID` provider. For the dense core, results are similar: 72.5% of stub ASes receive more than 60% of their traffic through their `lowID` link.

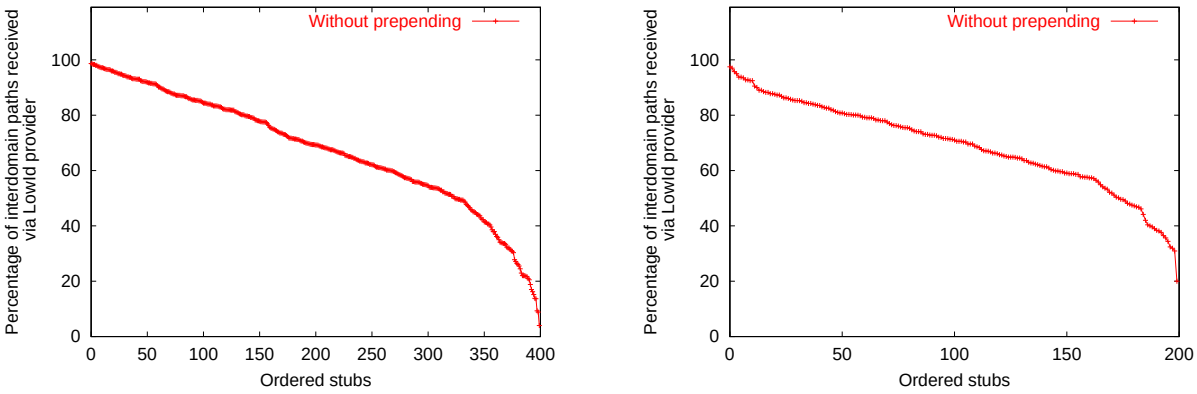

Fig. 12. Distribution of interdomain paths without prepending for lightly (left) connected and dense (right) core

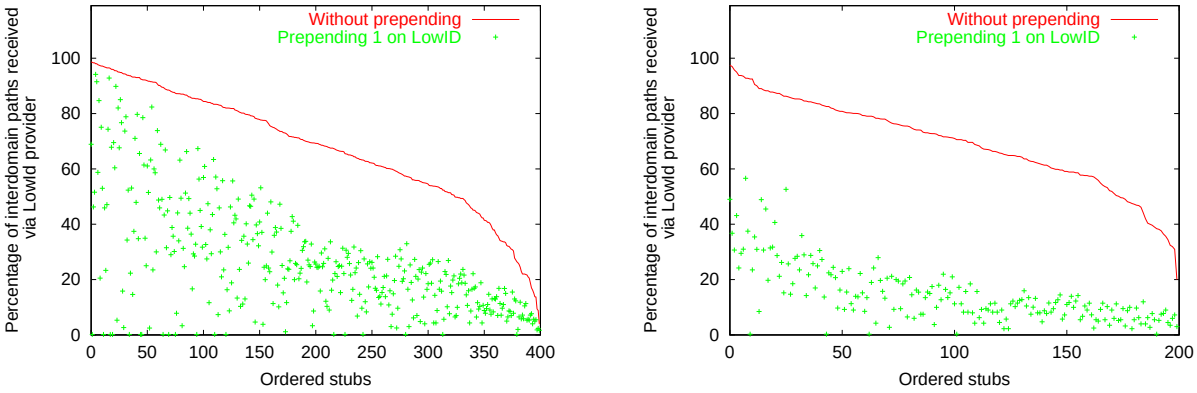

Fig. 13. Distribution of interdomain paths with prepending of 1 on the lowID link for lightly connected (left) and dense (right) cores

Impact of `AS-Path` prepending To evaluate the impact of `AS-Path` prepending we performed several simulations with the stub ASes configured to send prepended routes to one of their upstream providers. In all simulations, we configured all stub ASes in the same manner to ease comparisons. For our first simulation, each stub AS sent routes with its own AS number prepended once to its `lowID` provider and without prepending to its `highID` provider. In practice, a stub AS could use this prepending to better balance its traffic between its two upstream providers if it receives more traffic via its `lowID` provider.

Figure 13 shows the impact of this prepending on the distribution of the interdomain paths. In this figure, we show the distribution without prepending that was presented in figure 12 as a reference and use the ordering from this figure to plot the distribution of the interdomain paths with prepending.

The analysis of the simulation with the lightly connected core (figure 13, left) reveals several interesting results. First, as expected, the distribution of the interdomain paths is affected by the utilization of `AS-Path` prepending. One can see on figure 13 that with an `AS-Path` prepending of one on the `lowID` link, the distribution of the interdomain paths has changed for almost all stub ASes. With this amount of prepending, 79% of the stub ASes receive now less than 40% of their interdomain paths via their *lowID* provider.

However, a second important point to mention is that the influence of `AS-Path` prepending is different for each stub AS: some receive all their traffic through the *highID* link while other ASes seem not to be affected. This implies that it can be difficult for a stub AS to predict the impact of the utilization of `AS-Path` prepending on the distribution of its incoming traffic. This difference is due to the structure of the topology. In our topology as in the Internet, there exists a path between the two upstream providers of a stub AS. The length of the path between these two upstream providers determines the distribution of the interdomain paths after prepending. Let us first consider what happens when the two upstream providers of a stub AS are directly connected. In this case, the `lowID` provider will receive a direct route of two AS-hops and a route of two AS-hops through the `highID` provider. When comparing these two routes, the BGP decision process in our model relies on its random tie-breaker to select one over the other since no routing policies have been configured for *core* ASes. If the BGP decision process of the `lowID` provider selects the

route via the highID provider, then the stub will receive all the interdomain paths via its highID provider. When the two providers are not directly connected, then the impact of the distribution of the interdomain paths depends on their respective connectivity.
In the topology with the dense core, the utilization of AS-Path prepending has a stronger influence on the distribution of the interdomain paths as shown in figure 13 (right). After prepending, most stub ASes receive less than 40% of their interdomain paths via their lowID provider. As with the lightly connected *core*, after prepending some stub ASes do not receive traffic via their lowID provider anymore. The difference between the lightly and the dense core topologies can be explained by the connectivity of the providers. In the dense core, there are more direct links between providers and the providers with the lower connectivity have a much better connectivity than the less connected providers in the lightly connected core.

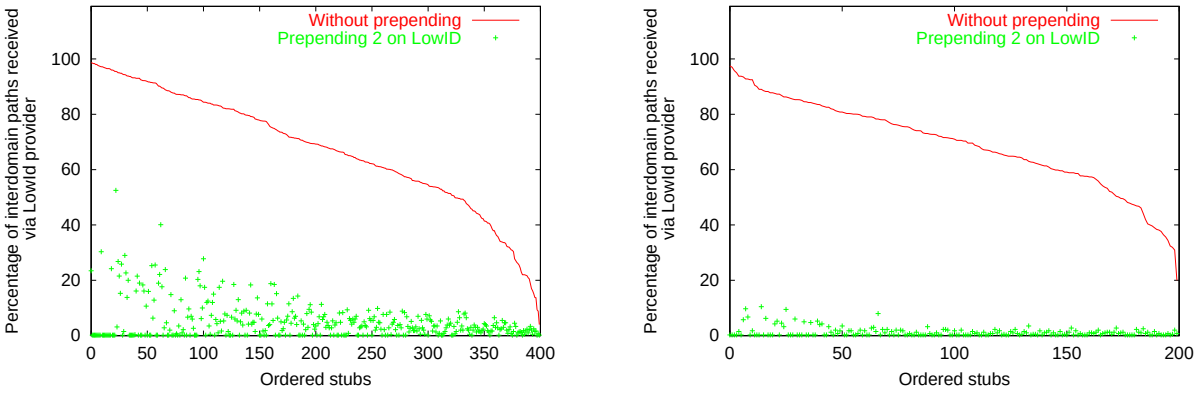

Fig. 14. Traffic distribution after prepending twice on the lowID link for the lightly connected (left) and the dense (right) cores

Prepending twice modifies significantly the distribution of the interdomain paths as shown by the simulation results in figure 14. For the lightly connected core, only 3 stub ASes still receive more than 30% of the interdomain paths via their lowID provider after prepending. Furthermore, 86% of the stub ASes receive less than 10% of their traffic through the prepended provider. This means that after prepending twice on the lowID link, almost all the interdomain paths have been shifted to the other link. The 3 stub ASes for which the effect is less important have actually a very good connectivity via their lowID provider and a very weak connectivity via their highID provider. On the dense *core*, prepending twice on the lowID link moves almost all the interdomain paths on the highID link.
Prepending 7 times, or more, is often used on backup links that should only be used in case of failures. Our simulations with this amount of prepending show that all the interdomain paths are received via the highID provider. This is because the topologies we used do not contain routes longer than 6 AS hops. This is similar to the current Internet, where most routes have a length between 2 and 4 AS hops and very few routes a longer than 6 AS hops.
We have also studied the effect of prepending on the link to the highID provider. Figure 15 shows that in this case, the distribution of the interdomain paths is much more affected than when prepending was used on the link to the lowID provider. This result was expected since the highID provider has a weaker interdomain connectivity than the lowID provider. As when prepending was used on the lowID link, the effect of prepending is not the same for all stub ASes. For the lightly connected *core*, 97 % of the stub ASes receive less than 30% of their traffic through their highID provider. A few stub ASes still receive a large part of their interdomain paths via their highID link provider despite the prepending. This is due to the good connectivity of the highID providers connected to these stub ASes.
For the dense core, the impact of AS-Path prepending on the distribution of the interdomain paths is even more important as shown in the right part of figure 15. Indeed, 84% of the stub ASes receive more than 95% of their interdomain traffic through their lowID provider.
Simulations with larger amounts of AS-Path prepending on the highID link have shown that most of the interdomain paths are received via the lowID provider. For example, for the dense core, all stub ASes receive less than 10 % of the interdomain paths via their highID provider when the stub ASes prepend twice the routes announced to this provider. For the dense core, all stub ASes receive less than 2% of the interdomain path via their highID provider after prepending twice.

Influence of local-pref In the previous section, we have studied the impact of AS-Path prepending by studying the distribution of the interdomain paths starting from each AS and ending inside each stub AS. This study

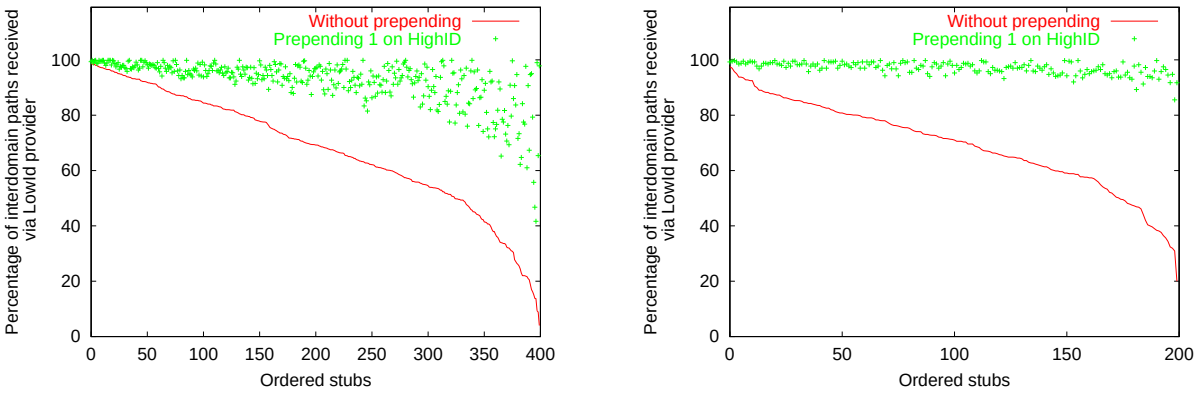

Fig. 15. Distribution of the interdomain paths after prepending once on the `highID` link for the lightly (left) and dense (right) cores

has been performed by assuming that each source of an interdomain path sends its packets along the best path selected by its decision process. However, as discussed in section 5, each AS can easily control its outgoing traffic by using the `local-pref` attribute which is evaluated first in the BGP decision process.

To evaluate the impact of the utilization of the `local-pref` attribute by the stub ASes, we performed the same simulations as above, but by first configuring `local-pref` on each stub AS to force it to send all its packets via its `lowID` provider. Figure 16 compares the distribution of the interdomain paths in this simulation with the distribution obtained (see figure 12) when each stub AS sent its packets along its best path toward each destination.

Surprisingly, based on this simulation result, the upstream provider selected by the stub ASes has only a minor influence on the distribution of the interdomain paths. This result is confirmed when we configured each stub AS to send their packets via their `highID` provider as shown in figure 16 (right). A closer look at the results shows that 68% of the stub ASes receive more than 60% of their interdomain paths through their `lowID` provider when each stub AS sends its packets via its `lowID` provider. On the other hand, 66% of the stub ASes receive more than 60% of the their interdomain paths through their `lowID` provider when each stub AS sends its packets via its `highID` provider. Similar results were obtained with the dense core and when prepending was used.

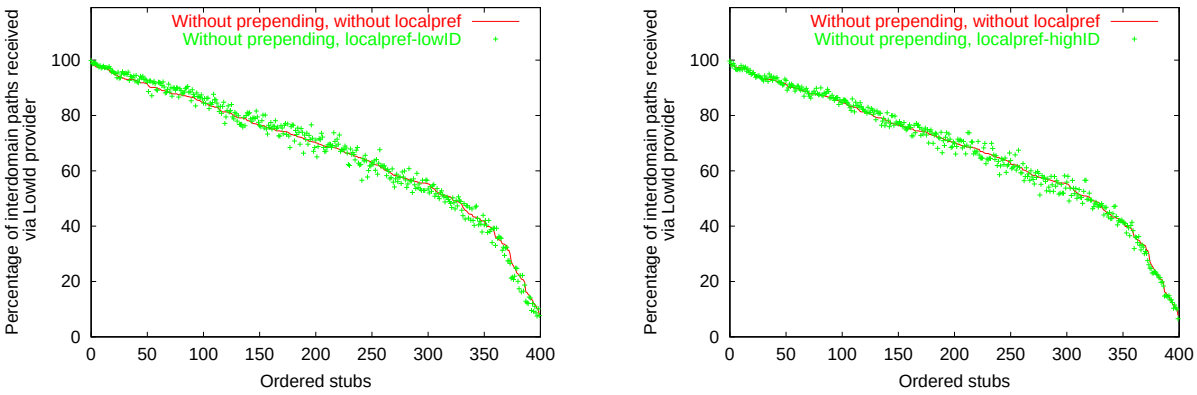

Fig. 16. Impact of `local-pref` on `lowID` link (left) and `highID` link (right)

This result can be explained by analyzing all the interdomain paths. A closer look at those paths reveals that a small number of *core* ASes appear in a large fraction of those paths. In figure 17, we show for the number of appearance of each *core* AS in these interdomain paths. Since each AS sent an IP packet with the `record route` option to each of the other 599 ASes reachable in our topology, there were 358801 interdomain paths. The analysis of those paths reveals that some ASes of the core appear in a very large number of interdomain paths and that many ASes only appear in a small number of paths. In the lightly connected core, one AS appeared in 100031 interdomain paths when the stub ASes sent their packets along their best path. This number changed to 109527 (resp. 105236) when the stub ASes sent

all their packets via `lowID` (resp. `highID`) provider. Figure 17 shows that the number of interdomain paths passing via each *core* AS does not change significantly when the stub ASes select one upstream provider or another.

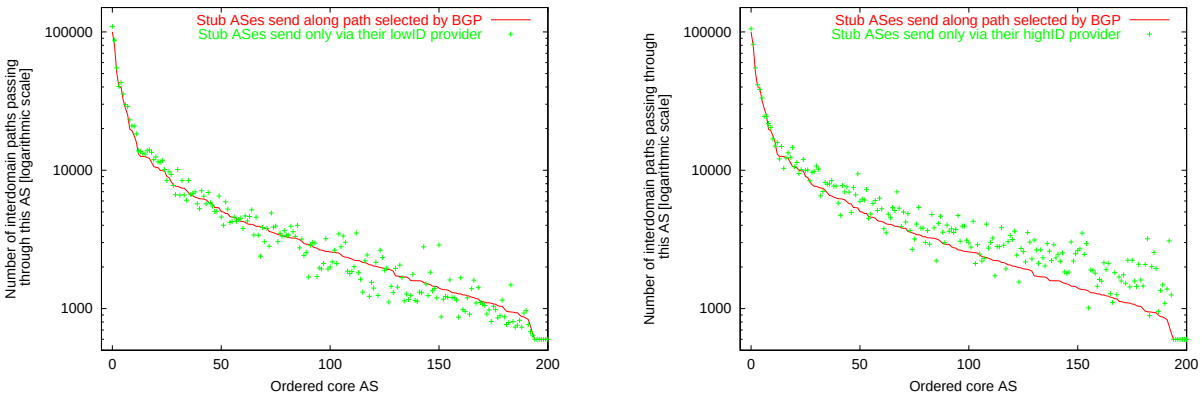

Fig. 17. Impact of the provider selected by the stub ASes on the interdomain paths

Based on these simulation results, it appears that the coupling between the traffic engineering techniques that allow stub ASes to control the flow of their incoming and the outgoing traffic is very weak. This weak coupling implies that the first hops of the path used by a stub AS to send its packets do not significantly influence the last hops of the path that these packets will take to reach their final destination.

7 Conclusion

In this paper, we have first described the behavior of BGP and explained several techniques that can be used to control the flow of interdomain traffic. We have also discussed the characteristics of interdomain traffic and have shown that although an AS will exchange packets with most of the Internet, only a small number of ASes are responsible for a large fraction of the interdomain traffic. This implies that an AS willing to engineer its interdomain could move a large amount of traffic by influencing a small number of distant ASes. Second, the sources or destinations of interdomain traffic are not direct peers, but they are only a few ASes hops away. This implies that interdomain traffic engineering solutions should be able to influence ASes a few hops beyond their upstream providers or direct peers.
We have then evaluated simulated the selection of two upstream providers by an hypothetical multi-homed ISP by relying on current BGP tables. Our analysis has shown several interesting results. First, when considering providers of similar size, they advertised routes of similar length. We have shown that on average, when an ISP has 2 Tier-1 upstream providers, both providers advertise routes with the same `AS-Path` length for about 60% of the routes in the BGP table. This implies that the length of the `AS-Path` is not anymore a sufficient criteria to rank BGP routes and that criteria to better engineer the outgoing traffic could be used easily.
We have then presented a detailed evaluation of techniques that can be used to control the flow of the incoming traffic. Our detailed simulations of `AS-Path` prepending has shown that it is difficult to utilize this technique to achieve a given goal. Our simulations with `local-pref` have shown that the utilization of this technique had only a small influence on the traffic received by remote stub ASes.

Acknowledgments

This work was partially supported by the European Commission within the IST ATRIUM project. This paper could not have been written without the BGP routing tables collected by Routeviews. We would also like to thank B.J. Presmore for his BGP implementation in Java and the authors of Javasim. We also thank Benoît Piret (YUCOM) and Marc Roger (Belnet) for the analyzed traffic traces.

References

[AB02] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Review of Modern Physics*, pages 47–97, January 2002.

[AEWX01] D. Awduche, A. Elmalid, I. Widjaja, and X. Xiao. A framework for Internet traffic engineering. Internet draft, draft-ietf-tewg-framework-05.txt, work in progress, May 2001.

[BA99] A.L. Barábasi and R. Albert. Emergence of Scaling in Random Networks. *Sciences*, (286):509–512, October 1999.

[Bar00] S. Bartholomew. The art of peering. *BT Technology Journal*, 18(3), July 2000.

[BBGR01] S. Bellovin, R. Bush, T. Griffin, and J. Rexford. Slowing routing table growth by filtering based on address allocation policies. preprint available from `http://www.research.att.com/~jrex`, June 2001.

[BNC02] A. Broido, E. Nemeth, and K. Claffy. Internet expansion, refinement and churn. *European Transactions on Telecommunications*, January 2002.

[Bor02] S. Borthick. Will route control change the internet ? *Business Communications Review*, September 2002.

[CBP93] K. Claffy, H. Braun, and G. Polyzos. Traffic characteristics of the T1 NSFNET backbone. In *INFOCOM93*, 1993.

[Cis99] Cisco. NetFlow services and applications. White paper, available from `http://www.cisco.com/warp/public/732/netflow`, 1999.

[CNO99] J. H. Cowie, D. M. Nicol, and T. Ogielski. Modeling the Global Internet. *Computing in Science & Engineering*, (1):42–50, Jan/Feb 1999.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *ACM SIGCOMM*, Sept. 1999.

[FP99] W. Fang and L. Peterson. Inter-AS traffic patterns and their implications. In *IEEE Global Internet Symposium*, December 1999.

[FRT02] B. Fortz, J. Rexford, and M. Thorup. Traffic engineering with traditional IP routing protocols. *IEEE Communications Magazine*, October 2002.

[GP01] T. Griffin and B. Presmore. An experimental analysis of BGP convergence time. In *ICNP 2001*, pages 53–61. IEEE Computer Society, November 2001.

[GW02] T. Griffin and G. Wilfong. Analysis of the MED oscillation problem in BGP. In *ICNP2002*, 2002.

[Hus01] G. Huston. Analyzing the Internet's BGP routing table. *Internet Protocol Journal*, 4(1), 2001.

[Ish] K. Ishiguro. Gnu zebra 0.92a. Available from `http://www.zebra.org`.

[KN74] L. Kleinrock and W. Naylor. On measured behavior of the ARPA network. In *AFIS Proceedings, 1974 National Computer Conference*, volume 43, pages 767–780. John Wiley & Sons, 1974.

[MAMB01] A. Medina, A.Lakhina, I. Matta, and J. Byers. BRITE: An Approach to Universal Topology Generation. In *MASCOTS 2001*, August 2001.

[McM99] P. McManus. A passive system for server selection within mirrored resource environments using as path length heuristics. Available from `http://www.gweep.net/~mcmanus/proximate.pdf`, April 1999.

[MGVK02] Z. M. Mao, R. Govindan, G. Varghese, and R. Katz. Route flap damping exacerbates internet routing convergence. In *ACM SIGCOMM'2002*, 2002.

[oO] University of Oregon. Route-views. Available from `http://antc.uoregon.edu/route-views`.

[PHS00] P. Pan, E. Hahne, and H. Schulzrinne. BGRP: A Tree-Based Aggregation Protocol for Inter-domain Reservations. *Journal of Communications and Networks*, 2(2), June 2000.

[Pre01] B. J. Presmore. Ssf implementations of bgp-4. available from `http://www.cs.dartmouth.edu/~beej/bgp/`, 2001.

[RL02] Y. Rekhter and T. Li. A border gateway protocol 4 (bgp-4). Internet draft, draft-ietf-idr-bgp4-17.txt, work in progress, May 2002.

[SARK02] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz. Characterizing the internet hierarchy from multiple vantage points. In *INFOCOM 2002*, June 2002.

[Ste99] J. Stewart. *BGP4 : interdomain routing in the Internet*. Addison Wesley, 1999.

[TGJ02] H. Tangmunarunkit, R. Govindan, and S. Jamin. Network Topology Generators: Degree-Based vs Structural. In *ACM SIGCOMM*, 2002.

[Tya02] Hung-Ying Tyan. *Design, Realization and Evaluation of a component-based compositional software architecture for network simulation*. PhD thesis, The Ohio State University, 2002.