A performance evaluation of BGP-based traffic engineering

Bruno Quoitin, Cristel Pelsser, Olivier Bonaventure*, Steve Uhlig
Computer Science and Engineering Department
Université Catholique de Louvain, Belgium
Email : (bqu,cpe,Bonaventure,suh)@info.ucl.ac.be
Phone : +32-10-479012, Fax: +32-10-450345

December 8, 2004

Abstract

Many Internet Service Providers tune the configuration of the Border Gateway Protocol on their routers to control their traffic. Content providers often need to control their outgoing traffic while access providers need to control their incoming traffic. We show, by means of measurements and simulations, that controlling the flow of the incoming interdomain traffic is a difficult problem. For this purpose, we first rely on detailed measurements to show the limitations of AS-Path prepending. Then, we show by using large-scale simulations that the difficulty of controlling the flow of the incoming traffic lies in the difficulty of predicting which BGP route will be selected by distant ASes.

1 Introduction

The Internet is composed of more than 16000 distinct domains operated by Internet Service Providers (ISPs), corporations or universities. These networks exchange reachability information by means of the Border Gateway Protocol (BGP) [1, 2]. BGP is thus an important building block of the Internet. However, it is a complex protocol which enforces various economical relationships among the domains. In the BGP terminology, a domain is often equivalent to an Autonomous System (AS).

There are two types of ASes in today's Internet. A stub AS is an AS that sends or receives IP packets, but does not transit packets. A transit AS is an AS that agrees to transit IP packets from one of its neighbors to another neighbor. [3] has identified two main types of relationships enforcing this classification. A *customer-to-provider* relationship is used when a customer AS buys connectivity from a provider AS. A *peer-to-peer* relationship is used when the connection cost is shared by the two ASes.

*Corresponding author

A stub AS is connected to one or several provider ASes that the stub uses as transit to send IP packets to any destination. The tendency for a stub AS is to be multiply-connected [4] to different providers for redundancy and Traffic Engineering (TE) purposes. Today, around 60% of customer ASes are multi-homed and this percentage is increasing [5].

For stub ASes, which represent the majority of the ASes today (82% according to [4]), controlling how the Internet traffic enters or leaves their network is an important problem. For instance, stub domains which provide content are interested in controlling the flow of their outgoing traffic. Indeed, they want to optimize the way information reaches their customers. Several techniques have been proposed to allow a content provider to optimize its outgoing interdomain traffic (see [6, 7] and the references therein).

On the other hand, access providers that serve small and medium enterprises, dialup or xDSL users will often want to control their incoming traffic because the customers they serve are mostly content consumers.

Interdomain traffic engineering covers the various techniques that enable ASs to control their interdomain traffic. Despite its importance, as explained in [8], interdomain traffic engineering is today more an art than a science. Furthermore, for many Service Providers (SPs), interdomain traffic engineering is still done on a trial and error basis. There have been very few detailed studies of the performance of those techniques.

In this paper, we focus on the control of the incoming traffic by multi-homed stub domains, a common operational problem. We rely on measurements as well as simulations to explain why this control of the flow of the incoming traffic is difficult.

The remainder of this paper is organized as follows. First, we summarize the main BGP traffic engineering techniques in section 2. Then, we present, in section 3, our measurement-based evaluation of the most common TE technique, AS-Path prepending. This is, to our knowledge, the first analysis of such measurements. Because measurements from one location cannot cover all possible scenarios, we then evaluate the BGP decision process by means of "Internet-scale" simulations. We first study the use of each rule of the BGP route selection process and show the importance of the tie-breaking rules 4.2. Then, we use simulations to evaluate the performance of AS-Path prepending in the global Internet. Finally, we present measurements with a second technique that relies on BGP communities.

2 BGP Traffic Engineering techniques

In this section, we briefly summarize the BGP-based traffic engineering techniques that can be used by stub ASes. A more detailed presentation may be found in [9, 10].

A key feature of BGP is the decision process used by each BGP router to select the best path to reach each destination, among all the received advertisements. In Figure 1 we see that the first rule compares the local preference of the routes. If multiple routes with the same local preference remain, the AS-Path length of the routes is compared. If a best route is obtained, the decision process ends. Otherwise, BGP relies on the subsequent rules to break the ties. These rules, called tie-breaking rules,

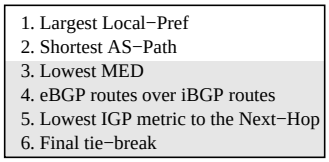

Figure 1: The BGP decision process.

are highlighted in Figure 1. The first tie-breaking rule checks the `MED` attribute. The routes which do not have the lowest value for this attribute are discarded. This attribute is used by a neighbor AS to influence the decision process in the local AS[9]. For instance, the `MED` can reflect the IGP cost of the path inside the downstream AS. The AS, thus, prefers the routes with smallest cost inside the downstream AS. Second, routes learned through eBGP sessions are preferred over routes learned through iBGP sessions. Routes learned through eBGP sessions are routes received from other ASes while routes received through iBGP sessions are routes received from internal routers. This rule thus implements the hot-potato routing principle since it will prefer routes that faster bring the traffic outside the AS. Then, the decision process compares the IGP distance to the next-hops of the remaining routes. The routes with the nearest next-hops are preferred. Finally, if more than one route remain, the BGP decision process relies on final tie-breaking rules which depend on the implementation. [1] recommends to tie-break on the `Router-ID` of the router through which we learned the route. However, some implementations have chosen to keep the oldest route [11] in order to avoid some oscillation problems.

Different mechanisms can be used to control the outgoing traffic and the traffic entering an AS. A stub domain only has to influence the decisions made by its own routers to engineer its outgoing traffic. For this purpose, it can easily prefer some routes over others by using the `local-pref` attribute for example. A common utilization of this attribute is to prefer routes learned from customers over routes learned from providers [12]. Other techniques rely on measurements to tune the outgoing traffic [13], [7].

However, if the stub domain is an access provider it usually has much more inbound traffic than outbound traffic. In this case, it often needs to control its incoming traffic, and the situation is far more complex. Indeed, the stub domain needs to influence the decisions made by routers in other domains. Several techniques exist [9]: selective announcements, more specific prefixes, AS-Path prepending, MED and redistribution communities. Unfortunately, announcing the prefixes selectively on peering sessions does not guarantee connectivity to the prefixes when a session fails. Moreover, certain ASes discourage to announce more specific prefixes, by dropping advertisements for small prefixes, in order to avoid an unnecessary growth of BGP routing tables [14]. The MED attribute should only be used when there are multiple physical links between two ASes and not in the case of stub ASes multi-homed to several providers, a very common situation today [5]. The only remaining techniques for multi-homed stub ASes are AS-Path prepending and BGP communities.

AS-Path prepending relies on the fact that the BGP decision process uses the length of the `AS-Path` to estimate the quality of a route. For this reason, a natural way to influence the choice of a neighbor router is to artificially increase the length of the `AS-Path` of certain routes to make them less preferable. Many network operators use AS-Path prepending on a backup line for instance or to deviate traffic from some neighbors without losing connectivity. A detailed analysis of BGP routing tables [15] revealed that 6.5% of routes were affected by prepending in November 2001.

Another technique that is becoming very popular is to rely on the BGP community attribute [16]. This attribute is an optional attribute that can be added to BGP routes. The presence of certain BGP communities inside a BGP route influence how this route will be processed by distant routers. Typically, an AS defines, in the configuration of its routers, a list of community values and the actions to perform when a route containing these community values is received. Customers of this AS may attach such communities to the routes they announce to this provider. The actions with TE purposes related to communities may be to request the provider not to announce the attached route to specific peers or to prepend the route on specific peering sessions.

Despite the importance of traffic engineering for ISPs, there have been few studies on the efficiency of these techniques. We are not aware of a detailed measurement-based evaluation of their performance. Moreover, the published simulation studies [17] were performed on topologies containing only a few hundred ASes.

3 AS-Path prepending measurements

In this section, we present a measurement-based evaluation of the performance of AS-Path prepending in a dual-homed stub AS. This study was conducted in the first semester of 2003.

3.1 Methodology

Our measurement study of AS-Path prepending was carried out by using a temporary stub AS number, AS2111, that was connected to two distinct providers: Belnet (AS2611), the Belgian Research Network, and a Belgian commercial ISP[1] (Be_ISP). At that time, Belnet was connected to 147 direct peers at several interconnection points and had two providers : TeliaNet (AS1299) and Level3 (AS3356). Belnet was also attached to the European research network, GEANT (AS20965). The Be_ISP was connected to 22 peers and had one provider. Figure 2 describes the AS-level topology as inferred from the BGP tables received from our two providers. This figure also shows the direct links between those providers. However, it should be noted that all these providers are also connected to larger ISPs.

To evaluate the performance of AS-Path prepending, we configured the BGP router of AS2111 to advertise a single `/20` prefix to its two providers. We did not perform measurements with more specific prefixes since as [14] we do not consider this as a suitable technique. Note that by using a `/20` prefix, we ensure that our prefix was not dropped by prefix length filters implemented by distant ISPs.

[1] We unfortunately cannot reveal the identity of this ISP.

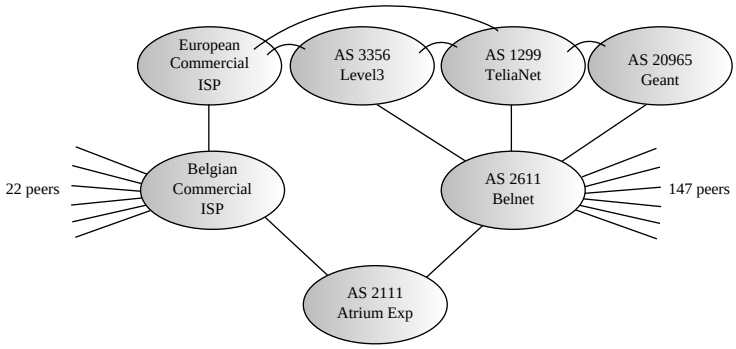

Figure 2: View from AS2111

Since our aim is to control the flow of the incoming traffic, the first step in our study consisted in generating traffic from as many sources as possible toward our prefix. Therefore, we gathered a list of valid IP addresses of `http` servers based on a one-month Netflow trace. We selected `http` server addresses because there were a wide variety of these inside the trace and we supposed that these addresses were persistent in time. We kept at most five addresses for each prefix in our BGP table. We completed this list of addresses by randomly selecting other IP addresses inside prefixes with less than 5 `http` servers inside the trace. We sent TCP `SYN` segments on port 80 to each address in the list, kept the addresses that answered and completed this new list with random addresses again. This way, we incrementally built a list of IP addresses responding on port 80 until we obtained at least one IP address responding for around 56000 prefixes out of the 125000 prefixes present in the BGP routing table. Moreover, the prefixes involved in our measures belong to 75% of the ASes present in the Internet at the time the measures were performed.

For each measurement, we modified the BGP configuration of the router in AS2111, restarted the BGP sessions with this new configuration and only started sending the TCP `SYN` segments two hours after the establishment of the BGP sessions to ensure the convergence of BGP for our route before the measurement. For each measure, we sent exactly one TCP `SYN` segment, with a given sequence number, to each of these destinations. We captured the responses on each interface by means of the `tcpdump` tool. This way, we determined, for each measure, the provider used by a given prefix to join our prefix.

3.2 Measurements

For our first measurement, we advertised our `/20` prefix without AS-Path prepending to our two providers. The first line of Table 1 shows that 67.82 % of the responses to the TCP `SYN` segments sent are received via Belnet. This difference is due to the variation in connectivity between our providers. Belnet is attached to two large ISPs and GEANT while Be_ISP is only attached to one large ISP. If we look at results at the

AS-level, 64% of the responding ASes sent all their replies via Belnet while only 27% of the replies arrived via Be_ISP.

A second interesting result is that for 9% of the ASes, we received replies via our two providers. Those replies came from different prefixes belonging to these ASes. This can be explained by two factors. First, most large ISPs use hot-potato routing to route the transit traffic. Consider for example a router of a Tier-1 ISP that peers with AS3356, AS1299 and the European Commercial ISP shown in Figure 2. When this router receives a packet whose destination is inside AS2111, it will send the packet to the closest router that is connected to one of the providers of our providers. Another router from the same Tier-1 ISP may select another transit AS to reach AS2111. This explains why, in our measurements, 9 of the 20 Tier-1 ISPs and 92 stub ASes advertising two or more prefixes but connected to a single provider, according to [4], sent replies via our two providers. Second, some stub ASes such as cable-modem providers are present in different cities. These ASes often use the cheapest provider in each city. They thus select different routes in different cities.

We then evaluated different amounts of AS-Path prepending via our two providers. Table 1 shows the impact of AS-Path prepending on each BGP session.

	Prepend to Belnet		Prepend to Be_ISP	
	Upstream		Upstream	
	Belnet (%)	Be_ISP (%)	Belnet (%)	Be_ISP (%)
no prepending	67.82	32.18	67.82	32.18
prepend once	22.22	77.78	79.64	20.36
prepend twice	15.67	84.33	80.87	19.13
prepend three times	15.35	84.65	100	0

Table 1: Impact of AS-Path prepending (prefix)

Table 1 shows that without prepending, the majority of prefixes reach AS 2111 via Belnet. When we look at the prepending of the AS-Path for our route advertisement on Belnet session (Table 1), we note that prepending our ASN once is enough to reverse the initial situation. Around 80% of the prefixes now respond via the Belgian ISP. Prepending the AS-Path twice still increases the ASes using the Belgian ISP to join our prefix. Additionally, we see that prepending the AS-Path three times on Belnet session doesn't have a significant impact compared to the distribution obtained by prepending the AS-Path twice on this session. Other prepending measures revealed that these prefixes could not be moved with an increased amount of prepending. It is very likely that local preferences influence the way their ASes send their traffic. For example, ASes connected to GEANT may favour GEANT for their outgoing traffic. This cannot be verified because it depends on the policies of distant ASs. However, as investigated later in this section, Figure 3 seems to confirm our assumption.

The same conclusions can be drawn for prepending on the link with the Belgian commercial ISP, in Table 1. Prepending the AS path once on the Belgian ISP session is enough to receive around 80% of the responses on the Belnet link. There is not much difference between prepending the AS path once and twice. This is due to the

length of the AS path in the advertisements received by the sources for our prefix. The selection of the best route towards AS 2111 depends on the policies enforced by other ASs as well as on the location of the Belgian ISP in the hierarchy of ASs. In this case 20% of the sources still prefer the path through the Belgian ISP even after this path is prepended twice. This is probably due to local preferences enforced by the European commercial provider for the customer routes of the Belgian ISP. Therefore, we see that it is required to prepend the AS-Path three times on the link to the Belgian ISP in order to switch all incoming traffic to the Belnet link.

To better understand the impact of AS-Path prepending, we selected from the BGP routing tables of our two providers, the list of their 10 most important upstreams. Those upstreams are the transit ASes located at two AS-hops from our providers that advertise routes for a large number of prefixes from which we receive replies via this provider. The largest of those ASes are mainly large Tier-1 ISPs.

Figure 3 shows the number of prefixes that we can reach via those ASes for which the replies are received via Belnet. For example, without prepending, Belnet received replies from more than 5000 prefixes whose routes received by Belnet contained `ASXX:AS701:*` in their AS-Path (where `ASXX` matches one of Belnet's providers). Note that since the Internet paths are asymmetrical, the AS-Path found in the BGP routing table of Belnet indicates the path used to send packets toward a given prefix, not the path used by those prefixes to reach our AS. Figure 4 provides the same information for the ten most important upstreams of Be_ISP. We see in Figures 3 and 4, that, for example, packets from prefixes that are reachable via AS701 (UUnet) and AS1239 (Sprint) are received via both Belnet and Be_ISP when no prepending is used.

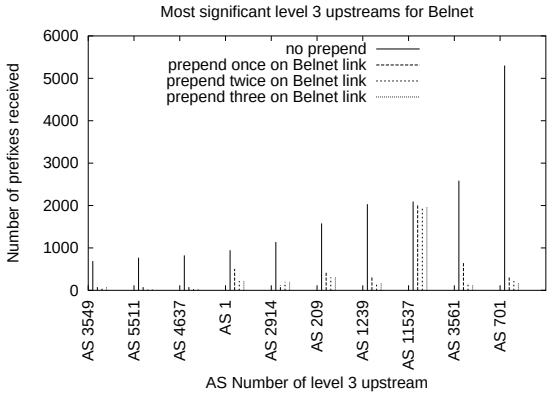

Figure 3: Important upstreams behind Belnet

When prepending is used on the Belnet link, Figure 3 shows that most of the prefixes move away from the Belnet link. However, a small fraction of the prefixes is not affected by the AS-Path prepending. AS11357 (Abilene) is special since almost none of the prefixes behind this AS move with AS-Path prepending. Abilene connects multiple universities and is connected to GEANT. For these universities, the Abilene connection is usually much cheaper than their commodity Internet connection and our

measurements indicate that they prefer their Abilene connection whenever possible.

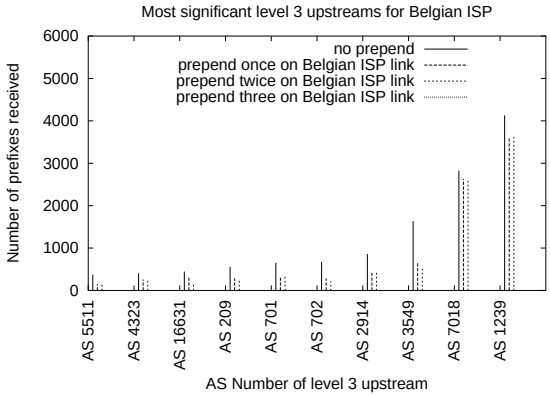

Figure 4: Important upstreams behind Be_ISP

Figure 4 shows that most of the responses from prefixes behind the 2 major upstreams do not move after prepending our announce on the Belgian ISP. However, over 50% of the responses from prefixes behind AS3549 (Globalcrossing) are received through Belnet when prepending is performed. When prepending three times is used on the link to the Be_ISP, all replies are received via the Belnet link. Note that those measurements correspond to the replies received from 56.000 prefixes. Some prefixes, not responding to our TCP `SYN` segments, may still send packets via the Be_ISP link even with prepending three times. To verify this, we used the BGP routing tables collected by RIPE [18] and Routeviews [19]. From the BGP tables collected at these sites, we were able to confirm our measurements. We also found one AS, AS11608 that did not respond to our TCP `SYN` segments and continued to use an AS-Path containing Be_ISP to reach AS2111 when the path via Be_ISP was prepended.

4 Simulation study of AS-Path prepending

The previous section has shown the limited efficiency of AS-Path prepending from a stub AS. However, measurement results may depend on the actual location of the stub AS and cannot easily be generalized to the whole Internet. In this section, we rely on simulations to study the efficiency of AS-Path prepending in the whole Internet.

4.1 A new BGP simulator

For the purpose of this paper, we developed a new and efficient open-source BGP simulator, *C-BGP* [20]. This was necessary because the other available open source simulators [21, 22] are unable to model networks as large as the Internet with 15.000 ASes. *C-BGP* is written in C and it has been used to perform simulations with more than 15.000 BGP routers.

In *C-BGP*, each BGP router is modeled as a data structure containing its RIB, Adj-RIB-IN and Adj-RIB-OUT. Each simulated BGP router is configured by specifying its physical interfaces, its eBGP and iBGP peers and the filters that are used on these sessions. *C-BGP* supports similar filters as those used on normal BGP routers [2]. *C-BGP* simulates the BGP messages that are used to advertise and withdraw prefixes over BGP sessions. These BGP messages can contain any valid BGP attribute. When a simulated BGP advertisement is received, this message is placed in the Adj-RIB-IN of the simulated router and the appropriate import filter is used. The BGP decision process is then run and a new BGP message is sent if a change in the best route occurred. For scalability reasons, *C-BGP* does not model the other BGP messages (OPEN, KEEPALIVE, ...), the underlying TCP connection and the various BGP timers (MRAI, HoldTimer, BGP dampening). Those mechanisms are important when evaluating transient issues such as the convergence of BGP but do not influence the selection of the best route with the standard BGP decision process [1].

To perform our simulations, we use an AS-level Internet topology that was inferred from real BGP routing tables gathered from multiple vantage points by Subramanian et al. [4]. The topology we used is dated from January 9th, 2003 and is the closest to our measurement period. It contains 14695 domains and 30815 interdomain links. There is at most one link between two different domains. We model each domain with a single BGP router and routing policies based on the economical relationships, determined by [4], which exist between the domains. To our knowledge, no simulation study has been able to analyze the impact of the routing policies on large networks composed of thousands of routers with routing policies. Most simulation studies only consider a few tens or sometimes a few hundred of routers. Given the importance of the routing policies, we choose to model them accurately. Memory constraints and the impossibility of inferring the internal topology of each AS from the available routing tables [4] forced us to consider a single router inside each AS.

The routing policy of our BGP routers is composed of two parts. The first part is the so-called *selective export rule* [3] which governs the provision of transit service. One domain provides a full transit service to its customers, a limited transit service between its customers and its peers but never between its providers and its peers. In our simulations, we configured each BGP router with the routing policies corresponding to the relationships with each of its peers. The second part of the policy introduces a preference among routes learned over different relations [3]. The routes learned from customers are preferred over routes learned from peers which in turn are preferred over routes learned from providers. The reason for such preferences is that providers do not have to pay their customers to carry traffic. This also ensures that interdomain routing will converge [12]. This policy is implemented in our simulations by relying on the `Local-Pref` attribute.

Although the January 9th topology is the most accurate publically available map of the global Internet, it has several limitations. First, in this topology each AS is modeled as a single node connected to neighboring ASes. In reality, an AS may contain up to several hundred of routers and there may be more than ten different physical links between two ASes although these links appear as a single edge in the inferred topology. Furthermore, the heuristic used to infer the routing policies is limited by two factors. First, it relies on a small set of BGP tables, typically collected at large

Tier-1 ISPs and those tables do not contain all interdomain links. For example, our measurements indicate that Belnet has 147 direct peers while according to the inferred topology, it only has 13 peers. This is normal since the January 9th topology was inferred on the basis of BGP tables from transit ASes. Second, the inferred routing policies are not always correct. For example, considering again Belnet, the inferred topology determined that Belnet has 9 providers, while as shown in Figure 2, it only has three providers.

4.2 Importance of the tie-breaking rules

Before analyzing the simulations with AS-Path prepending, it is important to understand how the BGP decision process selects the best path towards each destination.

For this purpose, we perform simulations with the model described in section 4.1. We instrumented the simulator to record, for each best route selected by a BGP router, the specific rule of the BGP decision process which was locally responsible for its selection. We then use this information to determine the importance of the different rules in the BGP route selection.

We perform 14695 simulations. In each simulation, a different domain announces a single prefix. We then count for each domain the number of routes selected by each rule of the decision process to join the announced prefix. For each source domain and each destination prefix, there are 4 possibilities. First, the domain has received a single route toward the prefix and the decision process is not applied. Second, the domain has received one route with a highest preference, thus the rule *local-pref* is accounted. Third, there are more than one route with the highest `Local-Pref` but one among them has a shorter `AS-Path`, thus the rule *shortest-path* is accounted. And finally, if there is still more than one route after the *shortest-path* rule, the *tie-break* is accounted.

We classify these results based on the level of the domain in the Internet hierarchy as identified by [4]. There are 5 levels in the Internet hierarchy. The first level is the core of the Internet and contains a full-mesh of large transit domains (tier-1's). The second level contains large transit networks (tier-2's) deeply interconnected. The third and following levels contain smaller regional providers and stub domains.

Figure 5 presents the results of these simulations. On the x-axis we show the 5 levels of the Internet hierarchy identified by [4]. The y-axis shows the relative importance of the rules, for each level of the hierarchy.x There is a bar for each considered rule: *local-pref*, *shortest-path* and *tie-break* as well as a bar for *single route*. The latter gives an idea of the importance of networks which only receive a single route to reach a destination and which thus do not apply the decision process.

The simulation results shown in Figure 5 reveal several interesting results. First, we can observe that between 30 and 45% of the Tier-3 to Tier-5 ASes only receive a single route to each destination. Those ASes are singled-homed ASes. For the Tier-3 to Tier-5 ASes, about 30 % of the BGP routes are selected on the basis of the length of their AS-Paths. The remaining 30 % of the routes are selected on the basis of the tie-breaking rules. For stub ASes, those tie-breaking rules often correspond to the `router id` step of the BGP decision process. Note that in reality, the `local-pref` rule may be used more frequently in stubs for backup or traffic engineering reasons, but this is not modelled in our simulations.

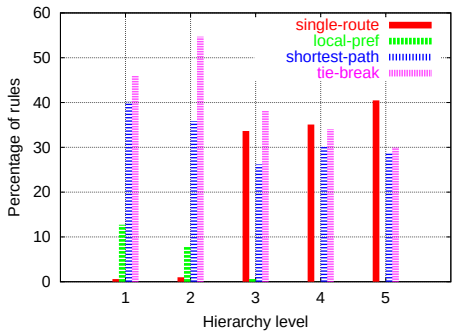

Figure 5: Importance of each rule of the BGP decision process at different levels of the Internet hierarchy.

Concerning the Tier-1 and Tier-2 ASes, 10% of the routes selected by those ASes are chosen on the basis of their `local-pref` attribute. Then, about 40 % of the routes are selected on the basis of their AS-Path. Finally, in the tier-1's and in the large national transit domains, about 50% of the routes are selected on a tie-break basis. This is due to the large number of interconnections that exist between all these domains and thus to a large number of alternative routes with a similar AS-Path length. For the transit domains, the tie-break rules correspond to the third, fourth and fifth step of the BGP decision process (Figure 1).

A consequence of the importance of the tie-breaking rules in the BGP decision process is that it is difficult to predict which best route will be selected in a remote AS. The selection of the best route depends on information that is not available outside the AS. Indeed, in the first tie-breaking rule, the value of the `MED` is only visible between neighboring ASes. In the second tie-break rule, if one eBGP route exists, the iBGP routes are removed from consideration. The outcome of the IGP metric rule depends on the internal IGP cost allocation policy of the considered domain. This information is usually confidential. Although some researchers have used `traceroute` to infer the IGP costs of internal links in transit ISPs [23], their accuracy appears to be limited [24]. In the final tie-breaking rule, we must know the Router-ID or the IP addresses of the involved routers, if the implementation relies on this information. Again, this is often kept secret by network operators. On the other hand, if the final tie-break keeps the oldest route, this decision is non-deterministic.

Since the tie-breaking rules are widely used in the BGP route selection, it is hard for an AS to evaluate how the traffic will enter the AS. Moreover, this also shows that the ASes often receive routes with the same AS-Path length for each destination prefix. We can already guess that this will influence the efficiency of AS-Path prepending. By increasing the length of the AS-Path for a route to one provider of a dual-homed stub, the route announced through the other provider is preferred by all ASes that used the tie-breaking rules for this destination. Consequently, a lot of incoming routes are likely to move to the preferred provider because the tie-break is used for above 30% of the routes.

4.3 Evaluation of AS-Path prepending

The aim of this section is to generalize our observations on the control of the routes entering our experimental AS with AS-Path prepending. Therefore, we perform simulations with the topology described in section 4.1, which captures a large portion of the real Internet. In the simulations we are not limited to a single dual-homed stub. We can obtain results similar to section 3.2 for each dual-homed stub in the topology. For each dual-homed stub, we study the use of `AS-Path` prepending to control how the other ASes reach the stub.

We rely on dual-homed stub domains to easily evaluate the impact of prepending on the distribution of the routes on their two upstream providers. These stubs represent 82% of the multi-homed stub ASes in the considered topology. The 5841 dual-homed stubs consist of more than 39% of the ASes in the January 9th topology. Single homed stubs are not considered since they do not have the possibility to engineer their traffic on multiple interdomain links. Stubs with more than 2 providers are less frequent. We do not consider them in this study because it is difficult to present graphically the simulation results for such multi-homed stubs.

We use the simulation model presented in section 4.1. For each considered stub, we determine how it is joined by all the other domains when no prepending is used. We then compute for each stub the distribution of paths via their two providers. We call it the "default" distribution. This distribution is plotted in Figure 6. To present the results graphically, we defined an ordering relationship among the providers. Each of our stubs has a well connected provider and a less connected provider. To determine what is the less connected provider of a dual-homed stub, we associate to each domain a ranking based on the following degrees: the number of providers of the domain, the number of peers and the number of customers. This ranking is a lexicographic order on ($\langle num\_prov,\ num\_peer,\ num\_cust \rangle$) to define the importance of a domain. This ordering has one exception for tier-1 domains in the core that do not have providers. When two domains have to be compared, if one is in the core and the other is not, the domain in the core is considered more important. Otherwise, the domain which has more providers is more important. If the number of providers is equal, we compare the number of peers, then the number of customers if required.

On the x-axis of Figure 6, we show the percentage of paths that cross the less connected provider. The y-axis of Figure 6 shows the number of stubs which have the same distribution of paths. We can observe that when no prepending is done, there is no clear tendency in favor of one of the two providers. Some stubs receive most of their interdomain paths via their most connected provider, others receive the same number of paths via each provider, while some stubs receive most of their paths through the less connected provider.

We then perform simulations where each dual-homed stub selectively prepends the `AS-Path` toward its less connected provider and toward its more connected provider. For each stub and for each of their providers, we use three different amounts of prepending: 1, 2 and 7. The results of these simulations are shown in Figure 7. The left plot shows the impact of prepending towards the less connected provider while the right plot shows the impact of prepending towards the most connected provider. The first important result that one can draw from these simulations is that the effect of prepend-

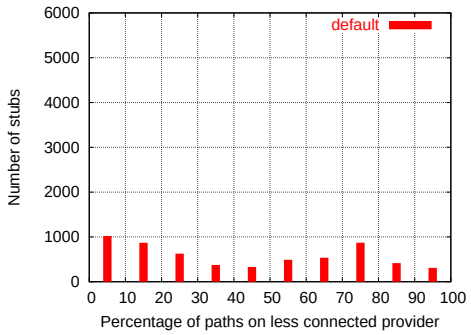

Figure 6: Default relative distribution of paths on the less connected provider.

ing is coarse. On average, prepending once toward one provider already moves a large fraction of the paths away from this provider. The granularity of `AS-Path` prepending is thus extremely limited. So is its interest for traffi c engineering.

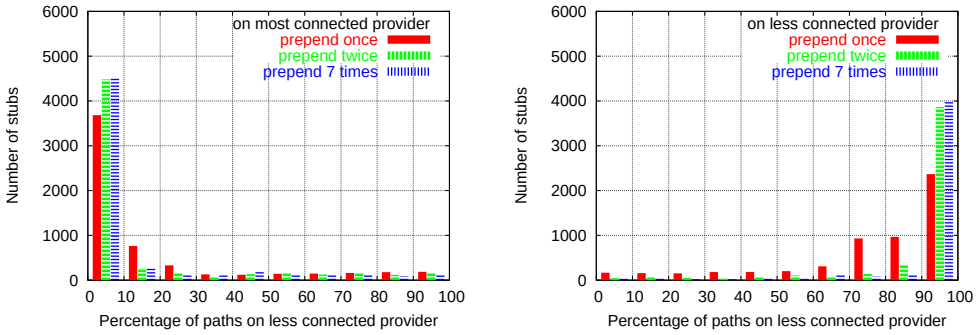

Figure 7: Percentage of paths on the less connected providers for various amount of prepending.

The second conclusion one can draw from these simulations is that the marginal benefi t of prepending decreases quickly. One can see that prepending once moves a lot of paths. Prepending twice still moves a lot of paths away. But the difference is minor between prepending twice and prepending 7 times. Furthermore, prepending too much can be a problem because inflated `AS-Paths` require an increased amount of memory in routers.

Third, the effi ciency of prepending is highly uncertain and depends on the location of the stub's providers as well as the relationships that these providers have with other domains. There are stubs for which paths can be moved easily from one provider to another, other stubs for which it is easier to move path from one provider to the other than the other way around and even stubs for which a very large part of the paths cannot be moved independently of the amount of prepending. Figure 8 presents three different stubs from the topology we have used and shows how the connectivity of their providers

constrain the efficiency of prepending. First, on Figure 8(a), the stub AS3748 has two providers, AS3786 and AS4766 which have a similar connectivity. They both have many *customer-to-provider* relations with domains in the core. The default distribution of incoming paths on the stub's access link is thus balanced: approximately 50% is received through each provider. This is due to the similar distance of the stub to the rest of the Internet through both providers. When prepending is used once toward AS3786, the percentage of paths which reach the stub through it decreases to 10%. This is explained by the distance of the stub which quickly becomes longer through provider AS3786 making the alternate path preferred. When prepending twice, this percentage falls to nearly 0%. The behavior is similar when prepending is used toward AS4766. After prepending once, the percentage of paths through AS4766 decreases to 3%. After prepending twice, it is close to 0%.

The second example, shown in Figure 8(b), shows a stub which has providers of different importances. The less connected provider, AS7066 has a single *customer-to-provider* link to AS1239 in the core. It also has a few customers. On the contrary, the second provider, AS7843 has three *customer-to-provider* relationships with AS1, AS209 and AS701, all in the Internet core. It also has two other relationships with minor providers and a few customers. Here, the default incoming path distribution is already unbalanced: 15% pass through AS7066 while 85% pass through AS7843. This is due to the choices of the domains in the core. They select the shortest path to the stub and re-advertise it to their clients and peers. In this case, the efficiency of AS-Path prepending differs when it is used toward the less connected provider or toward the more connected provider. After prepending once and twice toward AS7066, the percentage of incoming paths received through this provider becomes respectively 12% and nearly 0%. On the contrary, it is not possible to move all the incoming paths away from the other provider, AS7843. The results of prepending once, twice and 7 times give the following percentage of paths: 75%, 67% and 50%.

Another example is given in Figure 8(c). Here, the stub, AS17049 is also connected to two providers of different importance. The less connected provider is a priori AS6467 because it is not in the core while the other provider, AS1239 is. However, AS6467 has an excellent connectivity with domains in the core, such as AS1, AS701, AS7018 and also AS1239. Moreover, these domains (except AS1239) will prefer the routes learned from AS6467 which is a customer over the routes received from AS1239 which is a peer whatever the AS-Path length is! This is why after prepending only twice toward AS1239, there is already no more paths passing through it. On the contrary, prepending toward the other provider, AS6467, hardly moves a lot of paths. Even after prepending 7 times, the percentage of paths which reach the stub through AS6467 is still more than 58%.

5 Community-based traffic engineering

Besides AS-Path prepending, another technique that is often used to control the flow of the incoming traffic is to rely on BGP communities [16]. BGP communities are special values that are attached to BGP advertisements and used to request remote routers to perform some actions. The following traffic engineering actions are often supported:

- do not announce the route : in this case the route with the associated community should not be announced to the specified peers
- prepend n times when announcing the route : the AS-path of the route with the associated community will be prepended n times when it is announced to the specified peers
- specify the value of the local preference to be used by the router that receives the route [25].

These actions typically apply toward a large AS (e.g. tier-1 or tier-2 ISPs providing transit service), an interconnection point, a country or a continent. An extension to those BGP communities is currently being discussed within IETF [26, 27]. Unfortunately, all ISPs do not support all communities. For our measurements we had to rely on the communities supported by the providers of Belnet and Be_ISP.

Multiple community values can be attached to a route. However, in this section, we illustrate the influence of using single `do not announce` communities on the incoming traffic. In our measures, we first attached to the advertisement of our prefix toward the Belgian commercial ISP a community preventing the redistribution of our route by the European provider of Be_ISP to Sprint (AS 1239), then to AT&T (AS 7018) and, finally, to Globalcrossing (AS 3549). The results of these three measures are presented in Table 2. We note that a small portion of the ASes responding through Be_ISP reach our prefix through Belnet, when our prefix is not announced to Sprint by the European ISP. The same observation is made for the community preventing the redistribution of our route to AT&T by the European commercial ISP. However, the `do not announce` community toward Globalcrossing does not imply a move of the responses toward Belnet. Analogous results are obtained when using a community that requests the provider of Be_ISP to `prepend 3 times` its AS number when advertising our route to respectively AS1239, AS7018 and AS3549.

		Upstream	
	Belnet (%)	Be_ISP (%)	Both (%)
No communities	64	27	9
AS 1239	71	21	7
AS 7018	71	22	8
AS 3549	58	27	14

Table 2: Do not announce toward specified peer on the Be_ISP link

The classical BGP communities, as used in the measurements, or the redistribution communities being developed by the IETF [27] can be used to achieve a finer control on the incoming traffic. However, it should be noted that they suffer from three important drawbacks.

The first drawback is that, given our limited knowledge of the Internet topology and the routing policies used by distant ASes, it is difficult to predict the impact of a given community value. For example, consider Figure 2 and assume that AS2111

attaches to its route advertised to Belnet a community indicating that Belnet should not advertise the route to AS3356. In this case, AS3356 will not use its link with Belnet to reach AS2111. From Figure 2, AS3356 will send its packets to either AS1299 or the European Commercial ISP. In the first case, the community used by AS2111 does not have any effect on the packets received by AS2111. Furthermore, the sources that are downstream of AS3356 will recompute their best route to reach AS2111 and some of them may use AS1299 instead of AS3356 to reach AS2111 while others will utilize other paths. Given our limited knowledge of the Internet topology, it is very difficult to predict the decision that all those ASes will take.

A second drawback of the BGP communities is that the impact of one community on the incoming packet flow will depend on whether it is associated with other communities or not. For example, consider the right part of Figure 8. Assume that AS17049 uses a community to request AS6467 to not announce its route to AS701. In this case, AS701 may update its BGP routes and use its peering link with AS1 to reach AS17049 via AS6467. Thus, the community has no effect on the packet flow as seen by AS17049. However, if this community is used together with a community requesting AS6467 to not advertise the route to AS1, then both AS701 and AS1 will probably use AS1239 to reach AS17049.

Finally, the last drawback of the utilization of communities is that a typical AS will need to choose among a large number of different communities. For example, consider the redistribution communities [27] that allow a stub to influence the advertisement of its routes to the peers of its peers. The number of available redistribution communities depends on the number of ASes that are two AS-hops away. For Belnet, there are 1729 distinct ASes at two AS-hops.

In practice however, it can be expected that redistribution communities will mainly be used on customer-provider links. Figure 9 shows the cumulative distribution of the links at 2 hops for the multi-homed stub ASes in the topology from [4], on January 9th, 2003. This gives us an idea of the number of redistribution communities available at a multi-homed stub. The first curve (on the left) gives the cumulative number of multi-homed stubs with a given number of links with providers at 2 hops. The second curve concerns the number of peer-to-peer links at two hops. The third curve shows the number of links, at 2 hops, with customers that are single-homed. The last curves give the number of peerings with single and multi-homed customers at 2 hops and the total number of peerings at 2 hops.

Twenty percents of the multi-homed stubs have more that ten peerings with providers at two hops. These stubs can use $2^{10} = 1024$ sets of communities to engineer their incoming traffic with communities influencing the redistribution of their route toward providers only. Sixty percent of stubs have more than 500 links at 2 hops. This implies that a a lot of communities combinations exist to engineering the traffic of these stubs even if we exclude the communities targeting single-homed customers since this traffic cannot be moved with redistribution communities.

6 Conclusion

In today's Internet, ASes often need to control the flow of their interdomain traffic, for cost or performance reasons . In this paper, we have explained why it is difficult for an Autonomous System to control the flow of its incoming traffic.

We have presented a detailed measurement study of AS-Path prepending. Our measurements clearly show that the granularity of AS-Path prepending is very limited. In practice, this technique can be used to indicate that a backup link should be avoided whenever possible, but it is difficult to use it to balance the incoming traffic.

We have then used large-scale simulations to evaluate the BGP decision process and AS-Path prepending in the global Internet. An important finding of our simulations is that the tie-break rules of the BGP decision process are responsible for the selection of 30-50% of the routes in the global Internet. The simulations with AS-Path prepending confirmed the low granularity of this technique.

To accurately control the flow of its incoming packets, an AS should be able to predict which route will be selected by distant ASes. Unfortunately, this prediction is difficult for two reasons. First, our knowledge of the Internet topology and the routing policies is incomplete. Second, even with a detailed topology, it would still be very difficult to predict the outcome of the tie-break rules of the BGP decision process.

Based on our analysis, the current BGP-based techniques are not appropriate to control the incoming packet flow. Changes to the Internet architecture, such as those presented in [5] would probably be necessary to achieve such control.

Acknowledgements

This work was supported by European Commission within the IST ATRIUM and by the Waloon Government (DGTRE) within the TOTEM project. We thank Stefaan Decnodder and Sharad Agarwal for their review and helpful comments. We wish also to thank Jan Torreele from BELNET and the operators from the anonymous Belgian commercial ISP. Our measurements could not have been done without them.

References

[1] Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (BGP-4). Internet draft, draft-ietf-idr-bgp4-25.txt, work in progress, September 2005.

[2] B. Halabi. *Internet Routing Architectures (2nd Edition)*. Cisco Press, 2000.

[3] L. Gao. On inferring autonomous system relationships in the internet. In *Proc. IEEE Global Internet Symposium*, November 2000.

[4] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz. Characterizing the Internet Hierarchy from Multiple Vantage Points. In *INFOCOM 2002*, June 2002.

[5] S. Agarwal, C. Chuah, and R. Katz. OPCA: Robust interdomain policy routing and traffic control. In *IEEE Openarch*, New York, NY, April 2003.

[6] J. Bartlett. Optimizing multi-homed connections. *Business Communications Review*, 32(1):22–27, January 2002.

[7] S. Uhlig, O. Bonaventure, and B. Quoitin. Interdomain Traffic Engineering with minimal BGP Configurations. In *Proc. of the* 18^{th} *International Teletraffic Congress, Berlin*, September 2003.

[8] D. Awduche, A. Chui, A. Elwalid, I. Widjaja, and X. Xiao. Overview and principles of Internet Traffic Engineering. RFC 3272, May 2002.

[9] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, and O.Bonaventure. Interdomain Traffic Engineering with BGP. *IEEE Communications Magazine*, May 2003.

[10] Nick Feamster, Jay Borkenhagen, and Jennifer Rexford. Guidelines for interdomain traffic engineering. *SIGCOMM Comput. Commun. Rev.*, 33(5):19–30, 2003.

[11] E. Chen and S. R. Sangli. Avoid BGP Best Path Transition from One External to Another. Internet draft, draft-chen-bgp-avoid-transition-00.txt, work in progress, December 2003.

[12] L. Gao and J. Rexford. Stable internet routing without global coordination. In *SIGMETRICS*, 2000.

[13] D. Allen. NPN: Multihoming and route optimization: Finding the best way home. *Network Magazine*, February 2002. available from http://www.networkmagazine.com/article/NMG20020 206S0004.

[14] 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.

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

[16] O. Bonaventure and B. Quoitin. Common utilizations of the BGP community attribute. Internet draft, draft-bonaventure-bgp-communities-00.txt, work in progress, June 2003.

[17] O. Bonaventure, P. Trimintzios, G. Pavlou, B. Quoitin (Eds.), A. Azcorra, M. Bagnulo, P. Flegkas, A. Garcia-Martinez, P. Georgatsos, L. Georgiadis, C. Jacquenet, L. Swinnen, S. Tandel, and S. Uhlig. Internet Traffic Engineering. Chapter of COST263 final report, LNCS 2856, Springer-Verlag, September 2003.

[18] A. Antony and H. Uijterwaal. "routing information service - r.i.s. - design note". RIPE NCC document RIPE-200, http://www.ripe.net/ripe/docs/RIS/, October 1999.

[19] "University of Oregon Advanced Network Technology Center". University of oregon route views project. http://www.routeviews.org.

[20] B. Quoitin. C-BGP, an efficient BGP simulator. http://cbgp.info.ucl.ac.be, September 2004.

[21] B. J. Premore. SSF Implementations of BGP-4. available from http://www.cs.dartmouth.edu/~beej/bgp/, 2001.

[22] H. Tyan. *Design, realization and evaluation of a component-based compositional software architecture for network simulation*. PhD thesis, Ohio State University, 2002.

[23] Neil Spring, Ratul Mahajan, and David Wetherall. Measuring isp topologies with rocketfuel. In *Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 133–145. ACM Press, 2002.

[24] Renata Teixeira, Keith Marzullo, Stefan Savage, and Geoffrey M. Voelker. In search of path diversity in isp networks. In *Proceedings of the 2003 ACM SIGCOMM conference on Internet measurement*, pages 313–318. ACM Press, 2003.

[25] E. Chen and T. Bates. An application of the BGP community attribute in multi-home routing. RFC 1998, August 1996.

[26] A. Lange. Flexible BGP Communities. Internet draft, draft-lange-flexible-bgp-communities-02.txt, work in progress, March 2004.

[27] O. Bonaventure, S. De Cnodder, J. Haas, B. Quoitin, and R. White. Controlling the redistribution of BGP routes. Internet draft, draft-ietf-grow-bgp-redistribution-00.txt, work in progress, June 2003.

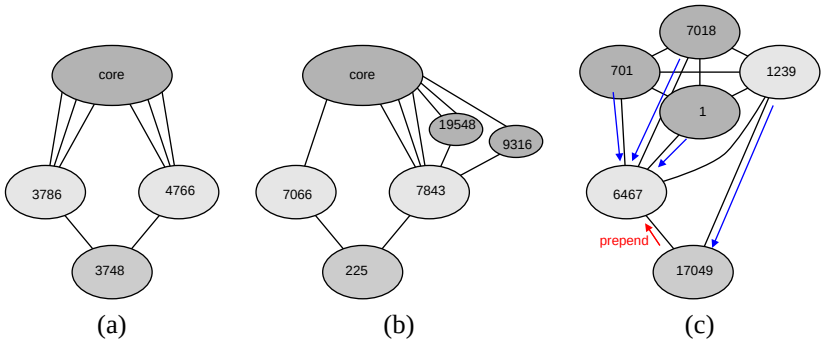

Figure 8: How the topology and economical relationships alter the prepending's efficiency.

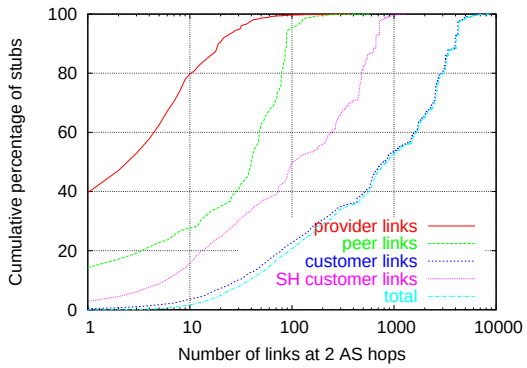

Figure 9: Importance of different business relationships at 2 AS hops from multi-homed stub domains.