Using redistribution communities for Interdomain Traffic Engineering*

B. QUOITIN, S. UHLIG and O. BONAVENTURE
(bqu,suh,obo)@infonet.fundp.ac.be

Infonet group
University of Namur, Belgium
http://www.infonet.fundp.ac.be

Abstract. In this paper, we present a traffic engineering technique that can be used by regional ISPs and customer networks. On the basis of various characterizations of ASes in today's Internet we show the requirements of the small ASes. Then we detail the methods that such ASes currently use to engineer interdomain traffic. We present an analysis of real routing tables showing that a lot of ISPs rely on the BGP Community attribute to build scalable traffic engineering configurations. We also show that this solution suffers from several drawbacks that limit its widespread utilization. To avoid the problems of such a technique, we propose the redistribution communities, a special type of non transitive extended community attribute and show that the cost of implementing this solution is small.

1 Introduction

Initially developed as a research network, the Internet has been optimized to provide a service where the network does its best to deliver packets to their destination. In the research Internet, connectivity was the most important issue. During the last years, we have seen a rapid growth and an increasing utilization of the Internet to carry business critical services such as e-commerce, Virtual Private Networks and Voice over IP. To efficiently support those services, several Internet Service Providers (ISP) rely on traffic engineering techniques to better control the flow of IP packets.

During the last years, several types of traffic engineering techniques have been developed [ACE+01]. Most of these techniques have been designed for large IP networks that need to optimize the flow of IP packets inside their internal network. These techniques are of very limited use for small IP networks that constitute most of the Internet today. For these networks, the costly resource that needs to be optimized is usually their interdomain connectivity. In this paper, we try to fill this gap by proposing a simple technique that can be used to provide useful traffic engineering capabilities targeted at, but not limited to, those small ISPs.

This document is organized as follows. We first discuss in section 2 the requirements for implementable interdomain traffic engineering techniques targeted at small ISPs. Then, we describe in section 3 the existing interdomain traffic engineering techniques. In section 4, we describe the redistribution communities that can be used to solve most of the traffic engineering needs of small ISPs.

* This work was supported by the European Commission within the IST ATRIUM project.

2 Interdomain traffic engineering for small ISPs

The Internet is currently composed of about 13.000 Autonomous Systems (AS) [Hus02] and its organization is more complex than the research Internet of the early nineties. Those 13.000 AS do not play an equal role in the global Internet. ASes can be distinguished on the basis of various characteristics like the connectivity one AS has with its peers, the services provided by one AS to its peers and the behaviour of the users inside the networks of one AS.

First, ASes can be distinguished on the basis of their connectivity. [SARK02] has shown that there are two major types of interconnections between distinct ASes: the *customer-provider* and the *peer-to-peer* relationships. The *customer-provider* relationship is used when a small AS purchases connectivity from a larger AS. In this case, the large AS agrees to forward the packets received from the small AS to any destination and it also agrees to receive traffic destined to the small AS. On the other hand, the *peer-to-peer* relationship is used between ASes of similar size. In this case, the two ASes exchange traffic on a shared cost basis. According to [SARK02], the *customer-provider* relationship is used for about 95 % of the AS interconnections in today's Internet.

Relying on this connectivity, [SARK02] makes a first characterization of ASes. There are basically two types of ASes: transit ASes that consitute the core of the Internet and regional ISPs or customer networks. The core corresponds to about 10 % of the ASes in the Internet and can be divided in three different subtypes (*dense*, *transit* and *outer core* depending on the connectivity of each AS). Regional ISPs and customer networks correspond to 90 % of the Internet and they maintain only a few *customer-provider* relationships with ASes in the core and some *peer-to-peer* relationships with other small ASes.

In this paper, we do not address the traffic engineering needs of ASes in the core but rather requirements of small ASes. The interested reader is referred to [FBR02] for a discussion of the needs of ASes in the core.

A second important element used to characterize an AS is the type of customer it serves. If the AS is mainly a content provider, it will want to optimize its outgoing traffic since it generates more traffic than it receives. On the other hand, if the AS serves a population of SMEs (Small and Medium Enterprises), dialup, xDSL or cable modems users, it will receive more traffic than it sends. Such ASes will typically only need to control their incoming traffic.

Another point to consider is the "topological distribution" of the interdomain traffic to be engineered [UB02a]. Although the Internet is composed of about 13.000 ASes, a given AS will not receive (resp. transmit) the same amount of traffic from (resp. towards) each external AS. The characteristics of the interdomain traffic seen from a customer AS have been analyzed in details in [UB02b]. In this paper, we have analysed the characteristics of all the interdomain traffic received by two small ISPs based on traces collected during one week. The first trace was collected at the interdomain routers of BELNET, an ISP providing access to universities and research labs in Belgium in December 1999. The second trace was collected during one week in April 2001 at the interdomain routers of YUCOM, a Belgian ISP providing a dialup access to the Internet. This study revealed two important findings that are summarized in figure 1. First, the left part of the figure shows the percentage of the number of IP addresses that are reachable

from the BGP routers of the studied ASes at a distance of *x* AS hops. This figure shows that for both studied ASes, most reachable IP addresses are ony a few AS hops away. Second, the right part of figure 1 shows the cumulative distribution of the traffic sent by each external AS during the studied week. The figure shows that for both ASes, a

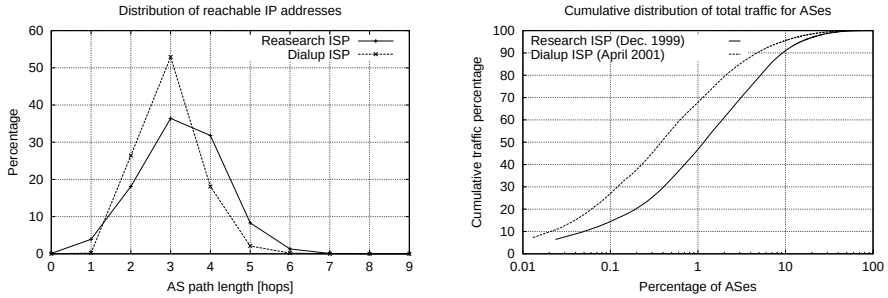

Fig. 1. BGP routing tables (left) and cumulative distribution of total traffic (right)

small percentage of external ASes contribute to a large fraction of the incoming traffic. Hence, by influencing this limited percentage of ASes a large fraction of the traffic can be engineered. Similar findings were reported in [FBR02] for an AS of the dense core.

3 Interdomain Traffic Engineering today

In this section, we review the traffic engineering techniques that are in use today in the global Internet. Since these techniques rely on a careful tuning of the BGP routing protocol, we first briefly review its operation.

3.1 Interdomain routing

The Border Gateway Protocol (BGP) [Ste99,RL02] is the current de facto standard interdomain routing protocol. BGP is a *path-vector protocol* that works by sending *route advertisements*. A route advertisement indicates the reachability of one IP network through the router that advertises it either because this network belongs to the same AS as this router or because this router has received from another AS a route advertisement for this network. Besides the reachable network, each route advertisement also contains attributes such as the `AS-Path` which is the list of all the transit ASes that must be used to reach the announced network.

A key feature of BGP is that it supports routing policies. That is, BGP allows a router to be selective in the route advertisements that it sends to neighbor BGP routers in remote AS. This is done by specifying on each BGP router a set of input and output filters for each peer.

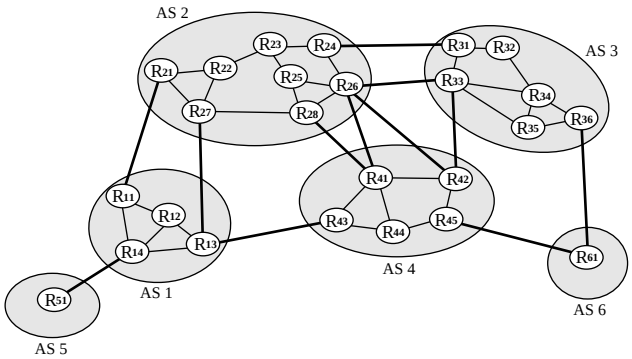

Fig. 2. A simple Internet

3.2 BGP-based traffic engineering

The BGP-based traffic engineering solutions in utilization today rely on a careful tuning of the BGP decision process [1] that is used to select the best-route towards each destination. This process is based on a set of criteria that act as filters among all the BGP routes known by the router.

Control of the outgoing traffic The control of the outgoing traffic is often a requirement for content providers that wish to optimize the distribution of their content. For this, they can rely on the weight and the `local-pref` attribute to control the routes that will be chosen for the packets that leave each BGP router of the content provider. The actual distribution of the outgoing traffic will depend on the quality of the setting of the weight and the `local-pref` on the BGP routers of the AS. The setting of these two parameters can be done manually based on the knowledge of the interdomain links or automatically with tools that rely on traffic measurements.

Control of the incoming traffic A customer AS serving a large number of individual users or small corporate networks will typically have a very assymetric interdomain traffic pattern with several times more incoming than outgoing traffic. These ASes typically need to optimize their incoming traffic only. For this, a first method that they can use is to announce different route advertisements on different links. For example in figure 2, if AS1 wanted to balance the traffic coming from AS2 over the links $R_{11} - R_{21}$ and $R_{13} - R_{27}$, then it could announce only its internal routes on the $R_{11} - R_{21}$ link and only the routes learned from AS5 on the $R_{13} - R_{27}$ link. Since AS2 would only learn about AS5 through router R_{27}, it would be forced to send the packets whose destination belongs to AS5 via router R_{27}.

[1] Due to space limitations, we cannot detail the BGP decision process in this paper. A description of the BGP decision process may be found in [FBR02,Hal97,QUPB02].

A variant of the selective advertisements is the advertisement of more specific prefixes. This advertisement 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). This fact can also be used to control the incoming traffic. In the following example, we assume that prefix `16.0.0.0/8` belongs to `AS3` and that several important servers are part of the `16.1.2.0/24` subnet. If `AS3` prefers to receive the packets towards its servers on the R_{24}-R_{31} link, then it would advertise both `16.0.0.0/8` and `16.1.2.0/24` on this link and only `16.0.0.0/8` on its other external links. An advantage of this solution is that if link R_{24}-R_{31} fails, then subnet `16.1.2.0/24` would still be reachable through the other links. However, an important drawback of advertising more specific prefixes is that it increases the number of BGP advertisements and thus the size of the BGP routing tables ([BNC02]).

Another method would be to allow an AS to indicate a ranking among the various route advertisements that it sends. Based on the BGP decision process, one possible way to introduce a ranking between routes to influence the selection of routes by a distant AS is to artificially increase the length of the `AS-Path` attribute. Coming back to our example, `AS1` would announce the routes learned from `AS5` on links $R_{11} - R_{21}$ and $R_{13} - R_{27}$, but would attach a longer `AS-Path` attribute (e.g. `AS1 AS1 AS1 AS5` instead of `AS1 AS5`) on the $R_{13} - R_{27}$ link. The required amount of prepending is often manually selected on a trial and error basis. The manipulation of the `AS-Path` attribute is often used in practice ([BNC02]). However, it should be noted that this technique is only useful if the ASes that we wish to influence do not rely on `local-pref` and `weight`.

Community-based traffic engineering In addition to these techniques, several ASes have been using the BGP Community attribute to encode various traffic engineering actions [QB02]. This attribute is often used to add markers to announced routes and to simplify the implementation of scalable routing policies on BGP routers. The community attribute is a transitive attribute that contains a set of community values, each value being encoded as a 32 bits field. Some community values are standardized (e.g. `NO_EXPORT`), but the Internet Assigned Numbers Authority (IANA) has assigned to each AS a block of 65536 community values. The community values are usually represented as `ASx:V` where `ASx` is the `AS` number to which the community belongs and `V` a value assigned by `ASx`. The community attribute is often used to encode the following traffic engineering actions [QB02]:

1. Do not announce the route to specified peer(s);
2. Prepend n times the AS-Path (where we have found values for n generally ranging from 1 to 3) when announcing the route to specified peer(s);
3. Set the `local-pref` value in the AS receiving the route [CB96];

In the first case, the community is attached to a route to indicate that this route should not be announced to a specified peer or at a specified interconnection point. For example, in figure 2, `AS4` could configure its routers to not announce to AS1 routes that contain the `4:1001` community. If `AS4` documents the utilization of this community

to its peers, AS6 could attach this value to the routes advertised on the R_{45}-R_{61} ink to ensure that it does not receive packets from AS1 on this link.

The second type of community is used to request the upstream AS to perform AS-Path prepending for the associated route. To understand the usefulness of such community values, let us consider again figure 2, and assume that AS6 receives a lot of traffic from AS1 and AS2 and that it would like to receive the packets from AS1 (resp. AS2) on the R_{45}-R_{61} (resp. R_{36}-R_{61}) link. AS6 cannot achieve this type of traffic distribution by performing prepending itself. However, this would be possible if AS4 could perform the prepending when announcing the AS6 routes to external peers. AS6 could thus advertise to AS4 its routes with the community 4:5202 (documented by AS4) that indicates that this route should be prepended two times when announced to AS2.

Finally, the third common type of community used for traffic engineering purposes is to set the local-pref in the upstream AS.

Our analysis of the RIPE whois database [QB02] provides more details on the utilization of the community attribute to request a peer to perform path prepending, to set the local-pref attribute and to not redistribute the route. The survey indicates that the specified peer is usually specified as an AS number, an interconnection point or a geographical region.

4 Redistribution communities

The community based traffic engineering solution described in the previous section has been deployed by at least twenty different ISPs, but it suffers from several important drawbacks that limit its widespread utilization. First, each AS can only define 65536 distinct community values. While in practice no AS today utilizes more than 65536 community values, this limited space forces each AS to define its own community values in an unstructured manner. [2] Second, each defined value must be manually encoded in the configurations of the BGP routers of the AS. Third, the AS must advertise the semantics of its own community values to external peers. Unfortunately, there is no standard method to advertise these community values. Some ASes define their communities as comments in their routing policies that are stored in the Internet Routing Registries. The RPSL language [AVG+99] used for these specifications does not currently allow to define the semantic of the community attribute values. Other ASes publish the required information on their web server or distribute it directly to their clients. This implies that an AS willing to utilize the traffic engineering communities defined by its upstream ASes needs to manually insert directives in the configurations of its BGP routers. Once inserted, these directives will need to be maintained and checked if the upstream AS decides for any reason to modify the semantics of some of its community values. This increases the complexity of the configuration of the BGP routers and is clearly not a desirable solution. A recent study has shown that human errors are already responsible

[2] We note however that facing the need for structured community values, some ASes like AS9057 have started to utilize community values outside their allocated space [QB02] and that other ASes are using community values reserved for standardization. This kind of selfish behavior is questionable in today's Internet, but it shows the operationnal need for more structured community values.

for many routing problems on the global Internet [MWA02]. An increasing utilization of community-based traffic engineering would probably cause even more errors.

A second drawback of the BGP community attribute is its transitivity. It implies that once a community value has been attached to a route, this community is distributed throughout the global Internet. To evaluate the impact of the communities on the growth of the BGP tables [Hus02], we have analyzed the BGP routing tables collected by RIPE RIS [RIS02] and the Route Views projects [Mey02][3] from January 2001 until now. A first observation of those BGP table dumps reveals that although most of the community values have a local semantics [QB02], a large number of community values appear in the BGP routing tables.

The evolution of the utilization of the communities reveals a sustained growth since the availability of the first dumps with community information in January 2001 (see figure 3). For instance, in recent dumps of routing tables provided by Route-Views

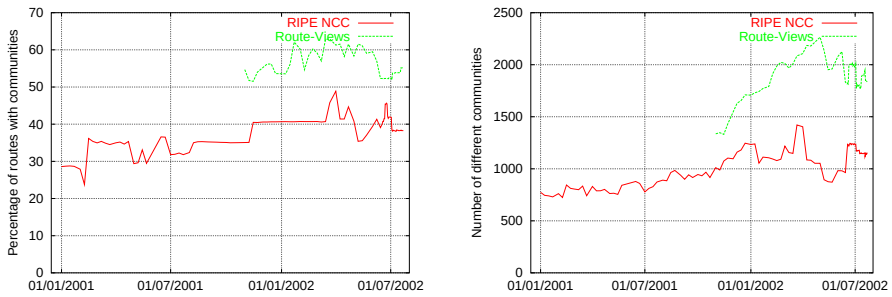

Fig. 3. Evolution of the utilization of the community attribute.

([Mey02]) at the beginning of the year 2002, the number of communities has increased to more than 2200 distinct values while more than 60% of the routes had at least one community attached and some routes can have up to 40 communities attached ! We could see the same evolution at other sites.

4.1 The redistribution communities

To avoid the problems caused by the utilization of the community attribute, we propose a new type of extended community attribute. The extended community attribute defined in [STR02] provides a more structured and larger space than the community attribute since each extended community value is encoded in an 8 octets field. The redistribution communities are non-transitive extended communities that can be used to encode a set of redistribution actions that are applicable to a set of BGP speakers. The current definition of the redistribution communities [BCH+02] supports the following actions:

– the attached route should not be announced to the specified BGP speakers.

[3] The Route-Views project started in November 2001

– the attached route should only be announced to the specified BGP speakers.
– the attached route should be announced with the NO_EXPORT attribute to the specified BGP speakers.
– the attached route should be prepended *n* times when announced to the specified BGP speakers.

Each redistribution community is encoded as an 8 octets field divided in three parts. The first octet is used to specify the type of non-transitive extended community [STR02]. The second octet is used to encode one of the four actions above and the last 6 octets encode a BGP_Speakers_Filter that determines the BGP speakers to which the action applies.

The BGP_Speakers_Filter field is used to specify the eBGP speakers that are affected by the specified action. There are two methods to specify the affected eBGP speakers. The first method is to explicitly list all those BGP speakers inside the BGP_Speakers_Filters field of redistribution communities. In this case, the high order bit of the BGP_Speakers_Filter field is set to 1. The second method is to explicitly list only the eBGP speakers that will not be affected by the specified action. In this case, the high order bit of the BGP_Speakers_Filter type field shall be set to 0. In the current specification [BCH+02], the BGP_Speakers_Filter can contain an AS number, two AS numbers or a CIDR prefix/length pair.

4.2 Implementation of the redistribution communities

In order to evaluate the cost of supporting the redistribution communities in a BGP router, we have modified the Zebra BGP implementation [Ish01]. The implementation of the redistribution communities requires two distinct functionnalities. The first one is to allow a network operator to specify the redistribution communities that must be attached to given routes and the second one is to influence the redistribution of the routes that have such communities attached.

First, in order to allow a network operator to attach redistribution communities to routes, we have extended the `route-map` statement available in the command-line interface (CLI) of Zebra. The `route-map` statement is an extremely powerful and versatile tool for route filtering and attribute manipulation that is composed of a filter and a list of actions. Our extension consists in the addition of a new action that can be used to attach a list of redistribution communities to routes that match the `route-map` filter. An example of a `route-map` using our new action is given below. The example presents the configuration in routers of AS6. This configuration attaches a redistribution community to every route announced to AS4. This community requests that AS4 prepend 2 times the AS-PATH of routes announced by AS6 when redistributing to AS2 (see example in section 3.2).

```
neighbor <as4-neighbor-ip> route-map prepend2_to_as2
route-map prepend2_to_as2 permit 10
  match ip address any
  set extcommunity red prepend(2):as(2)
```

Then, we have modified zebra so that redistribution communities are automatically taken into account. The implementation extracts the redistribution communities attached to the route and on the basis of their content, decides to attach the NO_EXPORT community, to prepend n times or to ignore the route when redistributing to specified peers. These modifications in the source code of Zebra were quite limited compared to the amount of work required to configure by hand redistribution policies similar to what redistribution communities provide. For instance, to establish the same configuration as shown above with a manual setup of communities, a lot more work is required.

5 Perspectives

Compared to the utilization of classical community values, the main advantages of the redistribution communities is that they are non-transitive and have a standardized semantics. The non-transitivity suppresses the risk of community-based pollution of routing tables while the standardized encoding allows simplification of the configurations of BGP routers and thus reduces the risk of errors [MWA02]. Furthermore, this will allow operators to provide services that go beyond the simple *customer-provider* and *peer-to-peer* policies currently found in today's Internet.

For example, BGP-based Virtual Private Networks [RR99] rely on communities to indicate where the VPN routes should be redistributed. The redistribution communities could be used to significantly reduce the configuration complexity of interdomain VPNs.

The redistribution communities could also be used to reduce the impact of denial of service attacks. For example, assume that in figure 2, AS6 suffers from an attack coming from sources located inside AS2. In order to reduce the impact of the attack, AS6 would like to stop announcing its routes towards AS2. With the standard BGP techniques, this is not possible while maintaining the connectivity towards the other ASes. With the redistribution communities, AS6 simply needs to tag its routes with a community indicating that they should not be redistributed towards AS2. If the attack originated from AS5, the redistribution communities would not allow AS6 to stop advertising its routes without also blocking traffic from sources like AS7. However, in this case, AS4 and AS1 might also detect the denial of service attack and could react with the redistribution communities.

In this paper, we have proposed a solution that allows an AS to influence the redistribution of its routes. It is nevertheless difficult to use a similar technique to influence the route redistribution farther than two AS hops away due to the variety and the complexity of the routing policies that might be found in the global Internet. However this is a first step towards a global interdomain level traffic engineering solution, which is our ultimate goal.

Acknowledgements

We would like to thank Russ White, Stefaan De Cnodder and Jeffrey Haas for their help in the development of the redistribution communities. The traffic traces were provided

by Benoit Piret and Marc Roger. We thank, RIPE for their whois database and their RIS project and Route Views for their routing tables.

References

[ACE+01] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and principles of internet traffic engineering. Internet draft, draft-ietf-tewg-principles-02.txt, work in progress, December 2001.

[AVG+99] C. Alaettinoglu, C. Villamizar, E. Gerich, D. Kessens, D. Meyer, T. Bates, D. Karrenberg, and M. Terpstra. Routing Policy Specification Language (RPSL). Internet Engineering Task Force, RFC2622, June 1999.

[BCH+02] O. Bonaventure, S. De Cnodder, J. Haas, B. Quoitin, and R. White. Controlling the redistribution of bgp routes. Internet draft, draft-ietf-ptomaine-bgp-redistribution-00.txt, work in progress, April 2002.

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

[CB96] E. Chen and T. Bates. An Application of the BGP Community Attribute in Multi-home Routing. Internet Engineering Task Force, RFC1998, August 1996.

[FBR02] N. Feamster, J. Borkenhagen, and J. Rexford. Controlling the impact of BGP policy changes on IP traffic. Toronto, June 2002. Presented at NANOG25.

[Hal97] B. Halabi. *Internet Routing Architectures*. Cisco Press, 1997.

[Hus02] G. Huston. AS1221 BGP table statistics. available from `http://www.telstra.net/ops/bgp/`, 2002.

[Ish01] K. Ishiguro. Gnu zebra – routing software. available from `http://www.zebra.org`, 2001.

[Mey02] D. Meyer. Route Views Archive project. University of Oregon, http://archive.routeviews.org, January 2002.

[MWA02] R. Mahajan, D. Wetherall, and T. Anderson. Understanding BGP misconfigurations. In *ACM SIGCOMM 2002*, August 2002.

[QB02] B. Quoitin and O. Bonaventure. A survey of the utilization of the bgp community attribute. Internet draft, draft-quoitin-bgp-comm-survey-00.txt, work in progress, March 2002.

[QUPB02] B. Quoitin, S. Uhlig, C. Pelsser, and O. Bonaventure. Internet traffic engineering techniques. Technical report, 2002. http://www.infonet.fundp.ac.be/doc/tr.

[RIS02] Routing Information Service project. Réseaux IP Européens, http://www.ripe.net/ripencc/pub-services/np/ris-index.html, January 2002.

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

[RR99] E. Rosen and Y. Rekhter. BGP/MPLS VPNs. Request for Comments 2547, Internet Engineering Task Force, March 1999.

[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.

[STR02] S. Sangli, D. Tappan, and Y. Rekhter. Bgp extended communities attribute. Internet draft, draft-ietf-idr-bgp-ext-communities-05.txt, work in progress, May 2002.

[UB02a] S. Uhlig and O. Bonaventure. Implications of interdomain traffic characteristics on traffic engineering. *European Transactions on Telecommunications*, January 2002.

[UB02b] S. Uhlig and O. Bonaventure. A study of the macroscopic behavior of internet traffic. under submission, available from `http://www.infonet.fundp.ac.be/doc/tr`, January 2002.