Explicitly Accommodating Origin Preference for Inter-Domain Traffic Engineering

Rolf Winter
NEC Labs Europe
Heidelberg/Germany
rolf.winter@neclab.eu

Iljitsch van Beijnum
Institute IMDEA Networks and UC3M
Leganés (Madrid)/Spain
iljitsch.vanbeijnum@imdea.org

ABSTRACT

Inter-domain traffic engineering is an important aspect of network operation both technically and economically. Traffic engineering the outbound direction is less problematic as routers under the control of the network operator are responsible for the way traffic leaves the network. The inbound direction is considerably harder as the way traffic enters a network is based on routing decisions in other networks. There are very few mechanisms available today that facilitate inter-domain inbound traffic engineering, such as prefix deaggregation, AS path prepending and systems based on BGP communities. These mechanisms have severe drawbacks such as an increase of the size of global routing table or providing only coarse-grained control. In this paper we propose and evaluate an alternative mechanism that does not increase the size of the global routing table, is easy to configure through a simple numeric value and provides a finer-grained control compared to existing mechanisms that also do not add additional prefixes to the global routing table.

Categories and Subject Descriptors

C.2.2 [**Network Protocols**]: Routing protocols

General Terms

Performance, Design, Standardization, Verification.

Keywords

BGP, traffic engineering.

1. INTRODUCTION

During the evolution of the Internet, Autonomous Systems (ASes) have become increasingly interconnected. Both at the edge and in the core of the Internet, ASes have continuously increased the number of other ASes they are direrctly connected to (see e.g. [1]). This trend is mainly driven by the need to increase both capacity and reliability of the connection to the global Internet. Given more than a single attachment point, an ISP can actually engineer its traffic, i.e., it can influence the way traffic leaves and enters its network.

To do that, ASes need to rely on the Border Gateway Protocol (BGP) which is used to exchange reachability information. BGP however is very limited when it comes to traffic engineering (TE). This is especially true for the inbound direction as the flow of traffic depends on the forwarding decision made at other routers in other Autonomous Systems. In other words, in order to engineer the way traffic enters a network, the route selection process at other routers has to be influenced remotely. Unfortunately, BGP has no obvious means built in that allows the origin of an advertisement to express a preference which could be used by other routers in the route selection process.

BGP is selecting a single best route towards a given destination, whereby a destination is represented by an IP prefix. If more than a single route is known to a BGP router, it follows an ordered sequence of steps to select the best amongst these routes. This sequence of steps is called the BGP decision process [2]. Each step in the process removes routes from the set of candidate routes until a single best route remains.

The BGP decision process has no mechanism directly built in that facilitates inbound TE. ISPs however have devised means to achieve their goal in three main ways. The first is to make the AS path longer by "prepending" their AS number multiple times. The second is making IP prefixes more specific and announcing them selectively – a practice called prefix deaggregation. Third, BGP community attributes can be used to perform a limited form of TE. Community attributes are not part of the decision process itself but can trigger certain actions to be applied to an advertisement such as the manipulation of path attributes or filtering.

Deaggregating an IP prefix into longer, more specific, prefixes is a fairly precise tool as all traffic will follow the more specific advertisement. However, it results in larger routing tables as more paths than necessary for reachability alone are injected into the global routing system. To put it differently, for the benefit of a single AS, all other ASes are burdened with additional routing table entries and the respective churn. AS path length manipulations on the other hand do not have the state issue. However, path prepending is a very coarse tool where a single AS prepend operation can result in dramatic traffic shifts [3]. This usually is too imprecise except for making one of the paths generally unattractive to select, e.g. a backup path.

Finally, community attributes are being widely used to give direct customers more control over the way their routes are distributed or handled by an upstream provider. E.g. communities could be used to control policy that is being applied to a route advertisement (e.g. setting the local preference in a certain, pre-defined range). Another typical application of communities e.g. is to perform a finer grained type of AS path prepending where an AS could define e.g. into which region (e.g. Europe) an advertisement should be prepended and how many times or even

SAC'12, March 25-29, 2012, Riva del Garda, Italy.
Copyright 2012 ACM 978-1-4503-0857-1/12/03…$10.00.

to which AS such an announcement should be propagated. A problem with communities is that they are widely used in non-standardized ways (as designed), i.e. based on local configurations. That means they cannot be globally interpreted and require configuration and planning.

It would be operationally beneficial to work with simple numerical values for preference that can simply be compared by routers in a standardized manner. In this paper we describe such a mechanism which hits the sweet spot between the precision of IP prefix deaggregation and the coarse control of AS path prepending without the configuration burden of communities and without increasing the global routing table size.

2. EXPRESSING ORIGIN PREFERENCE

With no appropriate TE means available, BGP needs to be extended to allow an origin AS to express its relative preference for a given prefix advertisement. Signaling origin preference is done through a new optional transitive path attribute which we call the Origin Preference Attribute (OPA, based on our work in [4]). It does what its name implies, it indicates an origin's relative preference for a given prefix announcement. The OPA is a 16 bit signed integer value which is set by the origin AS and is not changed by any intermediate AS. A higher OPA indicates a higher preference. To make the origin preference attribute an effective TE mechanism it needs to influence the route selection process.

2.1 Considering Origin Preference as part of the BGP Decision Process

Before we go into the details of how the OPA is influencing the decision process, we first explain where in the decision process the origin preference comparison should be placed. Generally speaking, the earlier in the decision process the origin preference is considered the more powerful this TE tool becomes and the more likely it will be used by origin ASes instead of e.g. prefix deaggregation. On the other hand, the earlier the OPA influences path selection, the less likely other ASes might want to implement it as it might interfere with their own local optimization goals. Given this intrinsic conflict, it is not easy to find the "right" place in the decision process.

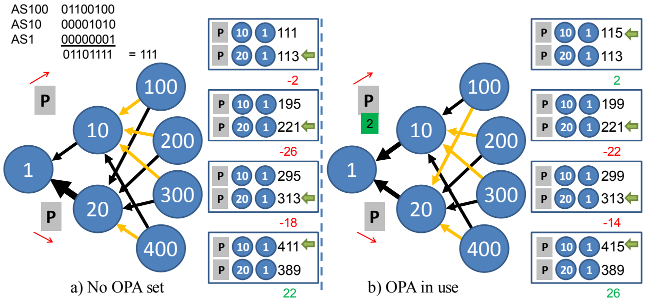
A good starting point for our deliberations are the two current practices that work on an Internet-wide scale to achieve inbound traffic engineering – AS path prepending and prefix desaggregation. Path prepending clearly influences the BGP decision process at the AS path length comparison step. Its ineffectiveness however is not due to the place in the decision process but its overly simplistic nature. Assuming local preferences all being equal, a single prepend results in all ASes at the same distance from the origin to change their path to lead through the shorter AS path. The simple numerical comparison of the AS path length and the dense AS level topology is what makes path prepending such a crude tool.

Prefix deaggregation is a very different mechanism and superficially, it seems that the decision process is not affected in the same manner as by AS path prepending as the same decision process is being executed just on smaller chunks of the same prefix. What effectively happens however is what RFC 3221 [5] calls "punching a connectivity policy 'hole'". This basically means that although an AS would prefer sending the traffic over a different AS, it is forced to choose another one as the origin has disaggregated the prefix and advertised it selectively.

The above leads us to the conclusion that today's inbound TE mechanisms already influence the decision process fairly early. We do believe that the AS path length comparison is important and we do not want to make path prepending ineffective as it is widely used. Therefore, we place the OPA comparison after the AS path length comparison in the decision process.

2.2 The OPA Decision Process

Using a simple numerical comparison of the OPA value will result in drastic traffic shifts, similar to AS path prepending. In order to have a mechanism that works more fine-grained, the OPA value is used in combination with another value that has properties somewhat comparable to a random number. More precisely, the OPA is added to this other value which we will refer to as random component R in the remainder of this document. R is calculated as follows. For each prefix advertisement received over EBGP, the origin AS number, the next hop AS number and the local AS number are all XORed to result in a 16 bit unsigned integer. For 32 bit AS numbers, the higher order and lower order 16 bits are simply XORed together. Take AS100 in Figure 1 as an example, where the origin AS (1) is XORer with the next hop AS number (e.g. 10) and the local AS number (100) to result in an R of 111 for the advertisement of P through AS10 and 113 for the advertisement of P through AS20.



a) No OPA set

b) OPA in use

Figure 1. The operation and effect of the OPA comparison

Each of the AS numbers that the R value is based on has an important role. The local AS number together with the next hop AS number create an R value that is (with some probability) different for every EBGP session. This will result in a different R value for the same prefix received from different next hop ASes. Important however is that the difference in these R values at one AS is different at other ASes at equal distance from the origin AS. To explain this, consider Figure 1. again. If the R values at AS200 were 211 and 213 (instead of 195 and 221) then the difference between these R values and the R values of AS100 were the same and the same OPA value would influence the path selection process equally. As can be seen in the figure, the R value differences are all different which is what enables a more fine-grained control using the OPA mechanism.

We need to apply a final modification to the operation of the OPA. R is a number between 0 and 65535. As the OPA itself is a singed 16 bit integer, there will be cases where the OPA is not sufficiently large to influence the path selection. We therefore multiply the OPA with two which results in an OPA range between -65536 and 65534. To cater for the corner case that the difference between R values is 65535 we cap the maximum R value to be 65533 so that the OPA can always influence the path selection process.

To follow our example in Figure 1 through. Assuming AS1 wanted to shift traffic from its link to AS20 onto the link between itself and AS10, it needs to add a positive origin preference value on its advertisement for P towards AS10 (or alternatively a negative value towards AS20). In Figure 1 b), AS1 adds a value of 2 to the advertisement towards AS10 which only results in AS100 to change its path selection. All other ASes keep the path they previously selected. Adding 10 in the example would have changed AS300's path selection in addition, and so forth.

There is one important additional constraint on the selection process. The OPA comparison only takes place in case prefix advertisements for a given prefix are received that have different OPA values. In case all advertisements have the same OPA value attached (or no OPA set), the set of candidate routes is not changed and the next decision process step takes place as the origin has no discernible preference. The subsequent tie breaking rules should be performed as they are important for the local AS, such as considering interior cost.

3. EXPERIMENTAL EVALUATION

We have implemented the OPA comparison in C-BGP [6], a BGP decision process solver, and simulated with it Internet-scale AS-level topologies based on data provided by UCLA [7]. The overall topology consists of over 30,000 ASes and from the same data set we use the inferred business relationships to set local preference values at the simulated routers. Our main goal was to evaluate the efficiency of the OPA-based inbound traffic engineering method on a reasonably realistic view of the current Internet. We used scenarios where the OPA comparison is globally deployed, i.e. every router includes the OPA comparison in the BGP decision process.

3.1 A. Behavior of Individual Prefixes – Dual-homed Case

In order to observe the effect that the OPA has, we started with simulations of a small set of stub ASes, i.e. ASes that do not provide a transit service for other ASes. At first, we chose ASes that are multi-homed to two different providers which constitutes the largest fraction of multi-homed ASes on the Internet. Towards one of the upstream providers we advertise a prefix with no OPA set, to the other provider we advertise the same prefix but with varying OPA values. To the former we will refer to as the non-OPA path/prefix, to the latter we will refer to as OPA path/prefix.

In Figure 2 we show five prefixes exemplary in the top part of the figure. The x-axis shows the OPA value on the OPA prefix and the y-axis shows the fraction of all ASes in the topology that pick the OPA path. Before starting the interpretation of the figure, an ideal mechanism would result in a figure that would be all straight lines, from the lower left side to the top right side. Practically this will of course not happen as the OPA comparison is not the first step in the decision process, R is not perfectly random, not all ASes actually receive both prefix advertisements, the decision at one AS can have a direct effect on the decision at other ASes and other reasons.

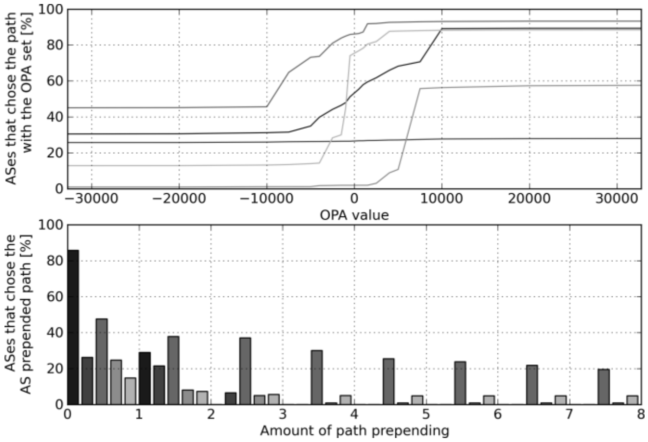


Figure 2. OPA comparison and AS path prepending compared

We picked those five prefixes because they nicely show that there is a certain range of potential outcomes (mainly depending on where in the topology the origin AS resides) when using the OPA and not because they are a good representation of the average case. There are a large number of prefixes for which the OPA mechanism works quite well. E.g. one prefix in the figure can move from about 10% of the ASes that pick the OPA path for the smallest possible OPA value to about 90% for the largest possible OPA value. In other words, when changing the OPA value from the most negative to the most positive value 80% of all ASes in the topology change their selected path to the OPA path. What the figure also shows is that for some ASes, the OPA does not really work. E.g one path is quite "unpopular" (around 30% of the ASes chose that path) although the OPA value is the largest possible. A smaller OPA value only minimally changes this over the whole range of OPA values. The same applies for paths on the other end of the spectrum (not shown), i.e. they are "popular" even with low OPA value.

The bottom half of the picture shows the same five prefixes for which we use AS path prepending to perform traffic engineering instead. The figure illustrates the rather coarse control of path prepending. Take e.g. the very left bar in the bar chart. A single path prepend results in well over 50% of all ASes to change their previously selected path. Another AS prepend results in virtually all ASes picking the shorter AS path. Using the OPA mechanism however, there is a much wider range of steps in-between.

The figure also shows that with AS path prepending there are various possible outcomes. In other words, AS path prepending already exhibit some of the behavior we see at the OPA decision process. Therefore, policy and the position within the topology, which determines how the prefix advertisements are propagated, are the likely factors that lead to this behavior.

3.2 Aggregate Behavior- Dual-homed Case

As a second step, we simulated 1000 dual-homed stub ASes to evaluate the behavior of a larger set of prefixes. Again, we let these ASes announce their prefix to one provider without any OPA set, towards the second provider varying OPA values were advertised.

Figure 3 shows the results of those experiments. In order to understand how often the OPA decision step is actually executed we first show in the upper left corner of the graph the average distribution of where in the decision process a router selects the route to install into the forwarding information base. About 48% only see a single route, mostly because they are single-homed. This is an important factor for the OPA mechanism because once a provider changes its decision because of the OPA decision process, all single-homed customers will also change at the same time. Additionally, a large number of routers pick a single best route at the path length comparison step and only about 7% on average stop at the OPA decision process. The number of routers that stop at the OPA decision process varies based on the OPA used. For small and large OPA values, the fraction of routers that stop at the OPA decision process is smaller than for small numerical values of the OPA (both positive and negative). The deviation from the 7% however (measured at an OPA of 1000) is only slightly above one percent. For such a small average number, the effect can be quite effective as we have seen before. Finally, still about 15 percent of the routers continue after the OPA decision process. This implies that these routers receive more than one route advertisement for a given prefix, however the OPA value that they observe does not differ.

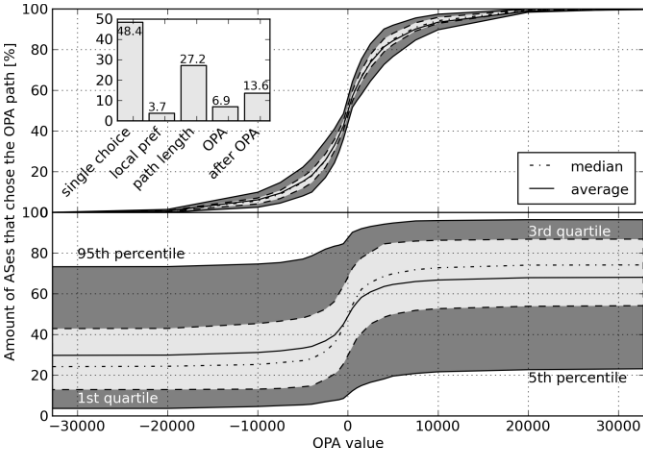


Figure 3. Aggregate behavior (dual-homed stub ASes)

Both, the top and bottom part of Figure 3 show essentially what the top part of Figure 2 has shown, i.e. the OPA value on the x-axis and the percentage of ASes that chose the OPA path on the y-axis. Here, the top part only shows the ASes that break at the OPA decision process, which illustrates the general behavior of the OPA mechanism. The bottom part in contrast shows the full aggregate view of the 1000 prefixes we simulated.

In the graphs, we plotted the 5th/95th percentile, the 1st/3rd quartile and the median and average. In the top graph, all these lines hardly deviate from each other, which shows that the decision process itself has a nice, even and predictable behavior. Also–as designed–when the OPA is low, none of these ASes pick the route with the low OPA, whereas when the OPA is high, all ASes pick the OPA path. For very large negative and very large positive OPA values there is hardly a change in the route selection process. Only starting at about -20000 and ending at about +20000 there is a noticeable effect on the decision process with the strongest effect (incline of the graph) when the OPA is a small positive or negative value. In this range of OPA values, there is also a small but noticeable increase in the amount of ASes that pick a route at the OPA decision process step as mentioned before.

The effect on all ASes, i.e. including the ones that do not pick the best route based on the OPA, is shown in the bottom part of the figure. As can be expected based on the behavior individual prefixes have shown before, the effect is not as pronounced and even. On average however, there is a significant amount of ASes that change their path selection based on the OPA mechanism. For the minimal OPA value, for 50% of all prefixes 25% or less of the ASes select the OPA path. At the other end of the spectrum around 75% or more pick the one with the OPA set. Given that only around 7% of all ASes pick a route at the OPA decision step; this is a very good outcome.

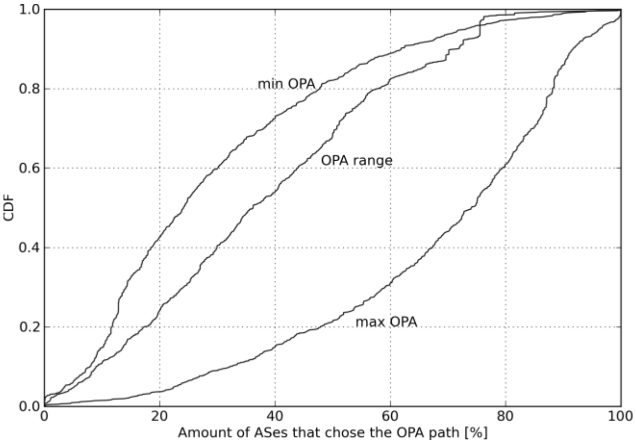


Figure 4. OPA range, upper and lower bound

Figure 4 shows a slightly different view of what happens. In the figure, we show the distribution of the percentage of ASes that pick the OPA path for the case where the OPA has the lowest and highest possible value. In addition we show what we termed the OPA range. With OPA range we mean the amount of ASes in percent that actually change their path selection when going from the minimal OPA value to the maximum OPA value, i.e the amount of ASes that can be influenced (directly and indirectly) using the OPA mechanism.

Each individual prefix is represented by a dot in the three lines ordered by the percentage of ASes that picked the OPA path. The top line shows the distribution for the minimum OPA value (lower bound). It illustrates how effective the OPA is to make a prefix announcement “unattractive”. Over 80% of the prefixes can be pushed below 50% of all ASes to pick OPA path. The bottom line shows the same for the maximum OPA value (upper bound). Here about 80% of the prefixes can be pushed above 50% of the ASes to pick the OPA path. An ideal mechanism would of course result in a line that already reaches 1.0 at an x-value of 0 for the minimum OPA case and a line that will stay 0.0 until an x-value of 100 is reached for the maximum OPA value.

The most interesting finding in the figure is however represented by the center line that shows the OPA range, i.e. the fraction of ASes that change their path selection based on the OPA from a minimum OPA value to the maximum OPA value. The graph shows e.g. that for 50% of the prefixes the fraction of ASes that can be influenced by the OPA is above 35%. The top 25% of the prefixes even have an OPA range of 55% and above. Again, the outcome employing the OPA mechanism is not perfect, but given the constraints, this represents a significant improvement over existing mechanisms.

4. Related Work

(Inbound) Traffic engineering is important but quite limited today. [8] is a good survey of traffic engineering techniques using BGP. Given its importance, it is not surprising that this field of work has received quite some attention in the past. Already early in the history of BGP version 4, attempts in standardization have been made to allow origin networks to have a larger degree of control over the way traffic enters their network, e.g. [9] and [10]. These proposals either suffered from a high degree of complexity or were severely underspecified. What it shows though that the need for such a mechanisms in neither purely academic nor new. It also shows that accommodating origin preference in BGP is not an easy task to accomplish. Other, more recent trends in the IETF show also that the current BGP decision process is too restrictive and operators would like to extend it, at least within their own domain [11]. Other efforts in this direction within the IETF have attempted to use e.g. well-known communities for the purpose traffic engineering [12].

There is also a large body of work from the research community. E.g. [13] suggests to rely on the existing configuration means available but let operators cooperate when changing configurations. Probably the most closely related piece of work is [14] which attempts to optimize AS path prepending. An algorithm is proposed that attempts to determine the optimal amount of prepending for a given prefix advertisement.

There is also lot of work that is complementary to our own work. E.g. [15] and [16] focuses on different aspects of egress path selection. Other work is more concerned with intra-domain traffic engineering yet other work considers mostly the outbound direction and general advice, e.g. [17].

5. Conclusion

Inter-domain traffic engineering is an important aspect of network operations today since a constantly increasing fraction of the networks that constitute the Internet is becoming multi-homed. Another trend that might need a tool different from prefix deaggregation is the recent depletion of the IANA IPv4 pool. This might result into smaller prefixes to appear in the global routing table over time as smaller allocations are made to customers. With e.g. only a /24 available, prefix deaggregation today will not work as many operators de facto filter on this boundary, i.e. a /25 or larger will likely not be globally routed. What this means is that such an AS is left with very little means to do efficient inbound traffic engineering.

The OPA mechanism that we have presented in this paper was an attempt to fill this perceived gap in BGP–a mechanism that can perform inbound traffic engineering finer-grained compared to the coarse control of path prepending but without the drawbacks of prefix deaggregation. We believe we have succeeded as the OPA mechanisms has proven to be quite effective for a large fraction of multi-homed stub ASes. We cannot claim to have devised the perfect inbound TE tool, but given the constraints of BGP and the nature of inter-domain routing, the OPA mechanism can significantly improve and nicely complement today's inbound TE tool set. Ultimately, the hope is to alleviate the need for deaggregation for a large fraction the prefixes observable in today's routing table which constitutes one third of the global routing table today–trend increasing. More research in this direction is however needed.

6. REFERENCES

[1] A. Dhamdhere, C. Dovrolis, "Ten Years in the Evolution of the Internet Ecosystem", In *Internet Measurement Conference (IMC)*, 2008.

[2] Y. Rekhter, T. Li, S. Hares, "*A Border Gateway Protocol 4 (BGP-4)*", RFC 4271, January 2006.

[3] B. Quoitin, C. Pelsser, O. Bonaventure, S. Uhlig, "A performance evaluation of BGP-based traffic engineering", *International Journal of Network Management* (Wiley), 2005.

[4] I. v. Beijnum, R. Winter, " A BGP Inter-AS Cost Attribute", *draft-van-beijnum-idr-iac-02* (work in progress), March 2009

[5] G. Houston, "*Commentary on Inter-Domain Routing in the Internet*", RFC 3221, 2001.

[6] B. Quoitin, S. Uhlig, "Modeling the routing of an Autonomous System with C-BGP", *IEEE Network*, Vol 19(6), 2005.

[7] *Internet Topology Collection*, http://irl.cs.ucla.edu/topology/.

[8] B. Quoitin, et al., "Interdomain traffic engineering with BGP", *IEEE Communications Magazine*, 2003.

[9] E. Chen, T. Bates, "Destination Preference Attribute for BGP", *draft-ietf-idr-bgp-dpa-05* (work in progress), September 1996.

[10] V. Antonov, "BGP AS Path Metrics", *draft-ietf-idr-bgp-metrics-00* (work in progress), March 1995.

[11] A. Retana, R. White, "BGP Custom Decision Process", *draft-retana-bgp-custom-decision-02*, (work in progress), May 2011.

[12] O. Bonaventure, et al., "Controlling the redistribution of BGP routes", *draft-ietf-ptomaine-bgp-redistribution-02* (work in progress), October 2003.

[13] J. Winick, S. Jamin, J. Rexford, "Traffic engineering between neighboring domains," *unpublished report*, July 2002.

[14] R. Gao, C, Dovrolis, E.W. Zegura "Interdomain Ingress Traffic Engineering through Optimized AS-Path Prepending", *IFIP Networking*, May 2005.

[15] A. Dhamdhere, C. Dovrolis "ISP and Egress Path Selection for Multihomed Networks", *IEEE Infocom* 2006.

[16] R. Teixeira, et al., "TIE Breaking: Tunable Interdomain Egress Selection", *IEEE/ACM Transactions on Networking*, August 2007.

[17] N. Feamster, J. Borkenhagen, J. Rexford, "Guidelines for Interdomain Traffic Engineering", ACM SIGCOMM Computer Communications Review, October 2003.