

The Border Gateway Protocol and its Convergence Properties

Ioana Kalaydjieva
(kalaydji@in.tum.de)

Seminar “Internet Routing”,
Technical University Munich,
June, 2003

Abstract

The Border Gateway Protocol plays a crucial role in the path establishing between autonomous systems in the Internet. It allows each autonomous system to independently define its routing policy with little or no global knowledge and to override distance metrics in favor of policy concerns when selecting a best route to a destination. Basing on an abstract BGP model, it will be shown that autonomous systems with mutually referential policies can result in non-terminating BGP route computation. A static approach towards this problem is to analyze different convergence properties in order to study the safety of routing policies. We will illustrate that the complexity of statically checking these properties is either NP-complete or NP-hard.

1.Introduction

1.1 The Structure of today's Internet

The Internet is divided into a large number of different regions of administrative control, commonly called *autonomous systems* (ASes). Examples range from college campuses and corporate networks to large Internet Service Providers. An AS has its own routers and routing policies and connects to other ASes to exchange traffic with remote hosts. There are two types of routing protocols to be mentioned in this relation – the IGP (Internal Gateway Protocols) and the EGP (External Gateway Protocols). The IGPs, also called *intra-domain* protocols, distribute local information within a single AS, while EGPs or the *inter-domain* protocols are responsible for the exchange of reachability information between different ASes. The only inter-domain protocol currently employed on the Internet is the Border Gateway Protocol (BGP). BGP allows ASes to apply local policies for selecting routes and propagating routing information, without revealing their policies or internal topology to others.

1.2 Routing Instability in Internet

The Internet engineering community is currently facing an important problem – the Internet routing instability, defined as the rapid change of network reachability and topological information. At the extreme Internet routing instability can lead to the loss of connectivity for large portions of the Internet. Having in mind the significance of Internet as a communication medium, it is easy to realize that such kinds of problems may nowadays generate millions of dollars of losses in

e-commerce revenue and interrupt the daily routine of hundreds of thousands of end users. So it seems extremely important to look for and analyze the possible origins of the Internet routing instability.

1.3 BGP Convergence Problem

Being interested in the investigation of routing anomalies, we should first pay special attention to the Border Gateway Protocol - the currently used inter-domain routing protocol. Typical of distance-vector protocols is the shortest-path routing, which means that if a choice has to be made between many routes to the same destination, the shortest one will be selected. However, BGP is not a pure distance vector protocol. It allows each AS to independently formulate its routing policies and it allows these policies to override distance metrics in favor of policy concerns, that is an AS may choose a non-minimal path because of its local policy. As the routing policies are implemented locally with little or no global knowledge, the interaction between even well-configured routing policies can result in some ASes exchanging BGP routing messages indefinitely, without ever converging on a set of stable routes, what is called route oscillation or protocol divergence. It has been proved that distance-vector protocols, because of the monotony of their distance metric, always converge. So, is it possible to guarantee the same about BGP? **That is what we call BGP convergence problem.**

The range of routing policies currently expressed, seem still limited, because the commercial Internet infrastructure is still in its infancy. However, it will grow both in size and topology complexity and conditions for route oscillations will become more and more likely to occur. That is why it is important to analyze such anomalies in order to avoid them or to recognize and get rid of them even if they should occur.

1.3.1. Two approaches towards the BGP convergence problem

The BGP convergence problem can be addressed either dynamically or statically. A dynamic solution represents a mechanism that suppresses or prevents at “run time” the BGP oscillations caused by policy conflicts. One such dynamic solution is the *route flap dampening*, which works in the following manner: The route flap dampening algorithms “hold down” (or refuse to believe) updates about routes that exceed certain parameters of instability, such as exceeding a certain number of flaps. A router will not advertise additional updates for a dampened route until a preset period of time has passed. However, route flap dampening does not eliminate the BGP oscillations, it just makes them run in slow motion. Another problem, connected with this solution, is that it is very difficult for network administrators to distinguish between route oscillations that arise from conflicting policies and others, caused by unstable routers or defective network interfaces i.e. network administrators are not supplied with enough information about the source of the route oscillations.

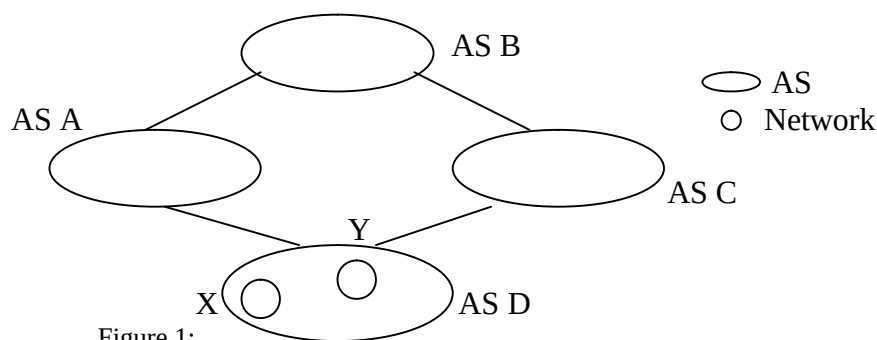
In contrast to a dynamic solution, a static solution relies on programs that check if the set of independently defined routing policies has any potential convergence problems i.e. whether the interaction of these routing policies could make the protocol to diverge. One good idea, which however turns out to be difficult to apply in practice for many reasons, is to design an architecture for routing policy coordination. In this architecture each autonomous system should first accurately **specify** its routing policy using a standard language (*RPSL – Routing Policy Specification Language*). Then, the autonomous system should globally publish that specification in a **routing registry**

(IRR's – Internet Routing Registries). The AS then should use this collection of routing policies through **analysis tools (RAToolSet)** to determinate connectivity or to detect conflicting policies. However, there are a number of practical difficulties, that such a solution to the BGP convergence problem faces and that make it hardly applicable in reality. On the one hand, it relies on voluntary publication of AS's routing policy. Many autonomous systems may be unwilling to reveal their local policies to others and may not keep the registry up-to-date. On the other hand, in the following sections of this work it will be shown, that even if ASes decide to reveal their local policies, **statically checking for convergence properties is either an NP-complete or an NP-hard problem**. These results suggest that the static approach to the BGP convergence problem may not be a practical one.

1.4 Paper outline

The remainder of this paper is organized as follows: Section 2 makes an overview on the functionality of the Border Gateway Protocol and introduces an abstract model of BGP, that is simpler than the real world BGP and allows to concentrate on the properties related to protocol convergence. In Section 3 there will be presented some examples of BGP systems, based on this model, that will illustrate the possibility of conflicting policies leading to protocol divergence (an unsolvable BGP system) or BGP systems having more than one solution. In Section 4 the complexity results concerning various properties of BGP systems will be discussed and it will be shown that the problem of statically checking them is NP-complete or NP-hard. The last Section 5 outlines the impact of these complexity results on the future researches of the BGP convergence problem.

2. The concept of BGP routing



An AS's traffic exchange agreements with its neighbors determine how the AS routes the traffic that enters or exits its infrastructure. For this reason, we say that AS's traffic agreements determine its *routing policy*.

The Border Gateway Protocol distributes the inter-AS routing information and is responsible for realizing routing policies. Conceptually, BGP works as follows: Each AS receives, from neighboring Ases, routes to destinations. In Figure 1 A tells B of a route to networks X and Y in D through itself. From C, B might hear of a route to the same networks. B selects one of those routes, causing its traffic to destinations in D to traverse that route. B then propagates this selection to its neighbors.

How can distribution of routes to destinations be used to realize routing policies? In BGP this happens in one of the two ways. An AS may *independently select* route

information received from its neighbors. For example *B* can, based on its routing policy choose the route through *C* to *D*, instead the route through *A*. In this manner, all traffic to *D* from *B* goes through *C*. An AS may also *selectively propagate* route information. For example, if *C* does not want *B*'s traffic to transit its infrastructure to reach *D*, it would not advertise to *B* that route to *D*.

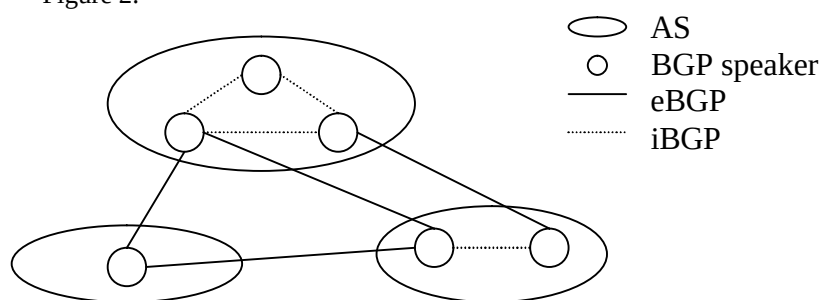
To sum it up, bilateral traffic agreements are implemented as routing policies that govern how routes to destinations are selected and distributed.

2.1 BGP in more details

Internet routing operates at the level of address blocks, or prefixes. A prefix may represent a single network, or a number of customer network addresses grouped into one larger “supernet” advertisement (address aggregation). Each prefix consists of a 32-bit address and a mask length; for example 192.0.2.0/24 consists of 256 addresses ranging from 192.0.2.0 to 192.0.2.255.

Routers in different ASes use BGP exchange update messages about how to reach different destination prefixes. In contrast to many interior protocols that build their own reliability on top of a datagram service, BGP uses TCP as its underlying transport mechanism. A router (BGP speaker) sends an *announcement* to notify its neighbor of a new route to the destination prefix and sends a *withdrawal* to revoke the route when it is no longer available. Each advertisement includes a number of attributes about the route, including the list of ASes along the path to the destination prefix. For this reason BGP is often referred to as a *path vector protocol*. Before accepting an advertisement, the receiving router checks for the presence of its own AS number in the AS path to detect and remove routing loops. An important property of BGP is that once advertised, routes do not need to be refreshed. They stay active until they are explicitly revoked or until the TCP connection breaks. When that happens, each router must stop using the information it learned from the other one.

Figure 2:



A router may receive routes for the destination prefix from multiple neighboring ASes. The router applies *import policies* to filter unwanted routes and to manipulate the attributes of the remaining routes. Ultimately, a router invokes *decision process* to select exactly one “best” route for each destination prefix among all the routes it hears. The router then applies *export policies* to manipulate the attributes and decide whether to advertise the route to neighboring ASes. In addition to exchanging BGP messages with neighboring domains, an AS may use internal BGP (iBGP) to distribute routing information amongst its routers (Figure 2). Ultimately, every router

must select a single best route for each prefix among the advertisements from the various eBGP (external BGP) and iBGP neighbors.

BGP advertisements can include numerous attributes like:

- ? **AS path:** Records the path of all autonomous systems that the route announcement has traversed
- ? **Next hop:** Contains the IP-address of the router that should be used as the next hop to the destination prefix
- ? **Local pref.:** The local preference attribute describes the sender's degree of preference for the advertised route. Each router has locally configured policies that determine this level of preference. This attribute is sent only to BGP routers within the same AS.
- ? **Origin type:** The origin type identifies how the origin AS learned about the route—within the AS (e.g. static configuration), EGP, or injection from another routing protocol. These origin types are known as IGP, EGP and INCOMPLETE.
- ? **Multiple exit discriminator:** A BGP advertisement may also include a multiple exit discriminator (MED) to encourage the recipient to pick a particular exit point for sending traffic to the neighboring AS.
- ? **Community:** The community attribute provides a generic mechanism for tagging routes to aid in specifying and applying routing policies. For example, an AS might assign different community values to a path depending on whether it was learned from a customer or a peer.

The BGP attributes are used by the *import* and *export policies* at each autonomous system in order to implement its *routing policies*. What about the *decision process* itself?

The selection of a best route to an AS is a function of both the router local policies and the best routes as determined by its neighbors. In other words, a router's choice of a route must be a one-hop extension of one of its neighbor's choices. A BGP-speaking router may learn multiple paths to the same destination prefix from eBGP and iBGP neighbors. First, certain routes are excluded from consideration.. This includes the routes removed during import processing (e.g. by a route filter or due to loop detection) and routes that have an unreachable next hop. Then, the router applies a sequence of steps to narrow the set of candidate routes to a single choice as follows:

1. **Highest local preference:** Prefer a route with the highest local preference, where local preference is assigned by the import policy and is conveyed via iBGP.
2. **Shortest AS path:** Prefer a route with the shortest AS path length¹.
3. **Lowest origin type:** Prefer a route with the lowest origin type (IGP is preferable to EGP which is preferable to INCOMPLETE)
4. **Lowest MED:** For routes with the same next hop AS, prefer a route with the smallest MED value, as conveyed in the BGP advertisement or reset by the import policy.

¹ Note that it is rule (1) that allows policy-based metrics to override distance-based metrics.

5. **eBGP over iBGP:** Prefer a route learned via eBGP over routes learned via iBGP, since leaving the AS directly is preferable to forwarding traffic through the AS to another router.

6. **Lowest IGP metric:** Prefer a route with the smallest intradomain (Interior Gateway Protocol) metric to reach the next hop, since this enables each router to select its closest exit point.

7. **Oldest route:** Prefer the route that was received earliest, since this route is more likely to be stable.

8. **Lowest router id:** Prefer the route learned from a router with the lowest router identifier, as conveyed during establishment of the BGP session. (Since several routers from the same AS can be running BGP at the same time, BGP identifiers are needed to distinguish between them.)

Even this brief overview on the functionality of BGP makes it clear that the real world BGP is rather complex for an investigation of properties related to protocol convergence. For this reason, the intent is to strip away all but the essentials from BGP and to design an abstract model in order to improve the clarity of the complexity statements and proofs.

2.3 An abstract BGP model

An abstract model makes it possible to study the safety of routing policies in a manner independent of the details used to implement those policies (for example, BGP attributes and import and export transformations). In modeling BGP, several simplifying assumptions, that do not affect the convergence properties, are made:

- 1) All issues relating to internal BGP (iBGP) are ignored
- 2) Address aggregation is ignored
- 3) It is assumed that there is at most one link between any two autonomous systems
- 4) It is assumed that no two autonomous systems can originate the same destination address
- 5) we assume that there is a default value for the local preference attribute: dlp

A BGP system is modeled as an undirected graph, where the nodes represent the different autonomous systems, and the edges – peering relations between them. During an *activation* each node (AS) applies its import and export policies and its path-selection process. When a node is activated, it can replace its current path with a path of a higher rank, if such path is available. It can also lose its path, if it is no longer available from its next hop, and be forced to accept a path of a lower rank (with lower local preference). Since each autonomous system operates independently, it cannot be assumed, that every AS is activated at the same time. Instead, we consider that a *subset* of all nodes is activated at a given time. The remaining ASes do not apply the path-selection process and hence, do not change their best route.

We will use BGP systems based on the abstract BGP model in order to analyze the BGP convergence properties. First we need to specify what is exactly meant by the convergence of BGP according to this model. Consider a BGP system of different

autonomous systems with individually configured routing policies. We start BGP sessions between some of the routers and assume that the topology does not change during the run of the protocol. There are **two possible behaviors of the system**:

- ? Eventually, all the routers reach the point, where they can not further improve their routes by extending the routes of their neighbors. When this happens, we say that the system has converged to a *stable* state or the system *has a solution*. That is, no matter what collection of ASes is activated, the system will remain in that state, because no autonomous system could pick a path better than its current path.
- ? Routers keep changing their preferable routes forever. In this case we say that the system *diverges* or *has no solution*.

The next section will introduce two examples of BGP systems – the unsolvable BAD GADGET and the BGP system DISAGREE, that has two solutions.

3. Examples of BGP Systems

3.1 BAD GADGET

In the absence of policy coordination it is possible for different autonomous systems to establish conflicting routing policies. More especially, three or more ASes may set up mutually dependent policies, that lead to non-terminating routing information exchange in BGP. Figure 3 shows a topology that can lead to such protocol behavior.

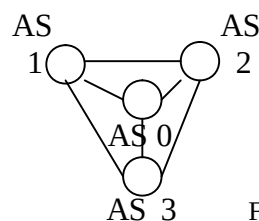


Figure 3: The BGP System BAD GADGET

Suppose, there is one single destination prefix d , originated by AS 0. The routing policies of AS 1, AS 2 and AS 3 are described as follows: Each AS prefers the counter-clockwise route of length 2 to all other routes to d .

According to our abstract BGP model that will mean for example for AS 3, that the route 3-2-0 will have the highest local preference, let say $dlp+1$ (where dlp is the default value for the local preference attribute), and will be preferred to other routes to the same destination like 3-0 and 3-1-0.

Now we will show that BAD GADGET has no solution. There are 16 spanning trees rooted at AS 0, but as the system is symmetric, we need only consider the six cases shown in Figure 4.

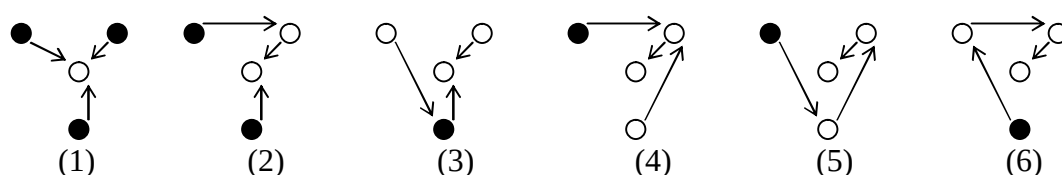
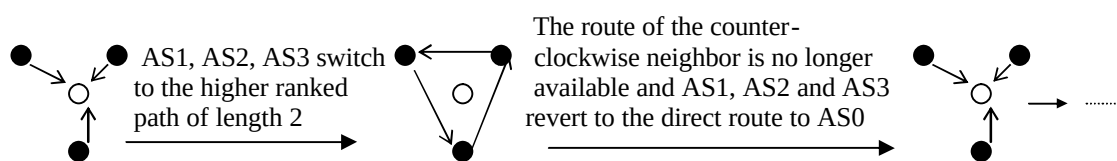


Figure 4: Possible routing trees for BAD GADGET

In this Figure the ASes that will change their selection of the best route to AS 0 are marked with a solid circle. Each marked AS will either pick the counter-clockwise route of length 2 (the route with the highest local preference), or revert to the direct route to AS 0.

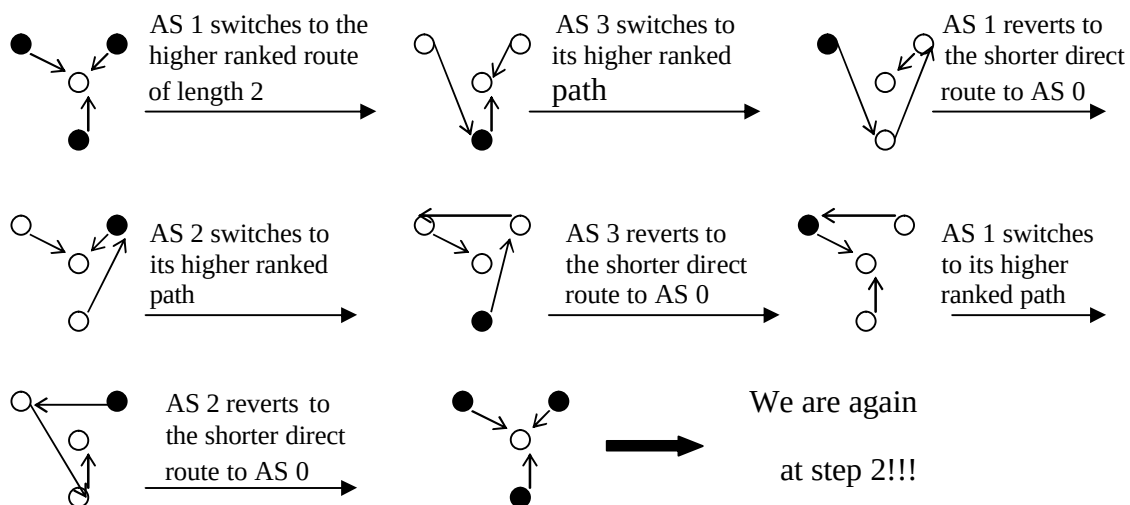
We will make an example with the first configuration of the possible routing trees (1), where AS 1, AS 2 and AS 3 have learned about a direct route to AS 0. Since we do not know whether the ASes are activated simultaneously, we will consider 3 different cases:

1. Simultaneous activation of AS 1, AS 2, AS 3:

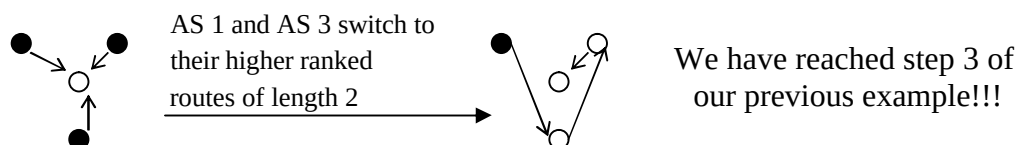


2. Non-simultaneous activation:

2.1 One autonomous system is activated (for example AS1)



2.2 Two of three ASes are activated simultaneously (for example AS 1 and AS 3)



In an analogous manner it can be shown for each of the six routing trees that BAD GADGET has no solution for the so defined routing policies, because they cannot be satisfied simultaneously.

3.2 DISAGREE

The system DISAGREE, presented in Figure 5, illustrates the fact that a BGP System can have multiple solutions and that a solvable system does not *always* converge on a solution.

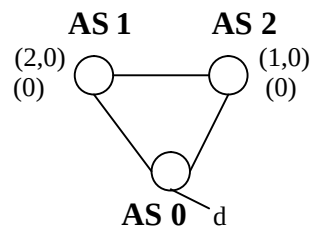


Figure 5: The BGP System DISAGREE

DISAGREE consists of three autonomous systems, that are connected pairwise. AS 0 originates a single destination prefix *d*. In Figure 5 with each AS is listed the set of possible routes in order of preference. Both AS 1 and AS2 prefer the path through the neighbor over the direct route to reach *d*. It is not difficult to see that DISAGREE has two stable states, whose corresponding routing trees are depicted in Figure 6.



Figure 6:

However, the system could also oscillate between two unstable states. In the first state, both ASes have selected the direct route (0). Then, if activated simultaneously, both ASes switch to their indirect routes (e.g., AS 2 switches to (1,0)). Then, if activated again, both ASes return to their direct routes, and the process is repeated. This sequence seems unlikely to arise in practice, because it relies on a precise sequence of events occurring repeatedly. If either AS 1 or AS 2 ever activates by itself, then the system will converge on a solution. The type of divergence possible with DISAGREE might be called *weak divergence* in contrast to the *strong divergence* of BAD GADGET, since it is possible to exit the “evaluation cycle” and arrive at a solution.

There are also further properties related to BGP convergence that could be analyzed with the help of BGP systems based on the abstract BGP model. For example it can be illustrated that even a solvable system can contain a trap leading to strong protocol divergence or that a single link failure is sufficient to make a solvable BGP System to diverge.

4. Complexity Results

Another BGP System – ASSIGN(n) – is designed in order to prove that statically checking BGP convergence properties like the solvability of a BGP System for instance, is an NP-complete problem. The proof relies on a reduction from 3-SAT, a well-known NP-complete problem. For a detailed presentation of the complexity results concerning various global properties of BGP systems and the corresponding proofs the reader is encouraged to consult [1].

5. Conclusions and Further Research

Even though the dynamic behavior of real-world BGP is considerably more complex than that of the abstract model, all of the results about the BGP convergence properties remain valid and provide *lower bounds* for the complexity of the corresponding questions for real-world BGP. So, since it turns out that the static analysis approach to the BGP convergence problem faces practical difficulties, efforts are made in finding other alternative solutions. Researchers are working on the possibility of restricting local policies in a way that guarantees BGP convergence, while still allowing greater flexibility than shortest-path routing [3]. Another idea is connected with an *extension* to the BGP protocol to carry additional information that would allow policy-based oscillations to be detected and identified at run-time [4]. For these reasons, an idealized version of BGP, called SPVP (Simple Path Vector Protocol) has been designed.

This paper gives just a vague idea of the complex problems in the Internet routing world. The problem of BGP divergence caused by routing policies interaction is only a tiny part of a big chain and many questions remain open. To which extent exactly is it a practical problem for today's Internet? And what about the future? Will the explosive growth of the Internet intensify BGP divergence to such an extent that revealing the proprietary routing policies will remain the only solution? Or will the extreme multiplication of the routes and routers lead to new, even more serious and acute problems that will push the BGP convergence problem into the background?

6. References

- [1] T.G.Griffin and G.Wilfong. An analysis of BGP convergence properties. In *Proc. ACM SIGCOMM*, September 1999.
- [2] T.G.Griffin, F.B.Shepherd and G.Wilfong. Policy disputes in path-vector protocols. In *Proc. Inter. Conf. on Network Protocols*, November, 1999.
- [3] L.Gao and J.Rexford. Stable Internet Routing without global coordination. In *Proc. ACM SIGMETRICS*, June 2000.
- [4] N.Feamster, J. Borkenhagen, J.Rexford. Controlling the Impact of BGP Policy Changes on IP Traffic. November 2001