BGP Route Propagation Between Neighboring Domains

Renata Teixeira, Steve Uhlig, Christophe Diot

BGP Route Propagation between Neighboring Domains

Renata Teixeira[1], Steve Uhlig[2], and Christophe Diot[3]

[1] Univ. Pierre et Marie Curie, LIP6-CNRS,
renata.teixeira@lip6.fr
[2] Delft University of Technology
S.P.W.G.Uhlig@ewi.tudelft.nl
[3] Thomson Paris Lab
christophe.diot@thomson.net

Abstract. We propose a methodology to match detailed BGP updates from two neighboring Autonomous Systems (ASes). This methodology allows us to characterize route propagation and measure the route propagation time. We apply this methodology to two months of all BGP updates from Abilene and GEANT to perform the first thorough characterization of BGP route propagation between two neighbor ASes. Our results show that the propagation time of BGP routing changes is very different depending on the network that initiates the routing change. This difference is due to engineering and connectivity issues such as the number of prefixes per BGP session, the number of BGP sessions per router, and BGP timer configurations.

1   Introduction

Although Autonomous Systems (ASes) in the Internet are independent management entities, events such as equipment failures or router misconfigurations in one AS can trigger BGP routing changes that propagate to other ASes. During routing convergence, user traffic may encounter loops or loss of reachability. Besides these transient disruptions, BGP routing changes can also lead to persistent reachability problems, because there may be no route to the destination or because the new route may be incorrect (in case of a misconfiguration). A detailed characterization of the dynamics of BGP route propagation can help reduce the impact of routing changes in one AS on neighboring ASes and reduce convergence delay. Such a characterization can also play an important role in diagnosing the root cause of persistent problems. To troubleshoot the problem operators often need to pinpoint the AS responsible for the routing change.

In this paper, we make a major step toward understanding BGP route propagation between neighboring ASes. We introduce a methodology for correlating BGP routing changes in two neighboring networks based on BGP updates collected in each of the ASes. We use this methodology, together with two months of BGP updates from Abilene and GEANT (the research backbones in the U.S. and Europe, respectively) to analyze BGP route-propagation time. Our results show that although the types of BGP routing changes that propagate between these two networks are similar, the propagation time is significantly different depending on which of the two networks initiates the routing change. We show how this disparity is based on each network's design and engineering decisions, including factors such as the number of prefixes per BGP session, the number of BGP sessions per router, and the configuration of BGP timers.

This is the first time that BGP update measurements from every router in two neighboring ASes have been used to evaluate the impact of BGP routing changes on neighbors. Previous studies of BGP dynamics either analyzed BGP update messages from multiple routers in the same AS [1–4] or a single router in each of multiple ASes, as available from RouteViews or RIPE, combined with beacon updates [5–7]. Analyzing BGP updates in one AS can reveal how routing changes propagate within a single network, but does not shed light on how these changes affect neighboring domains. Studies of multiple ASes can characterize the BGP convergence process in the wide area, without shedding light on the effects of intra-AS topology and configuration. In this paper, we find that per-router BGP measurements and knowledge of the network design and configuration details are essential for understanding the factors that affect route-propagation time.

The remainder of the paper is structured as follows. Section 2 presents background on BGP routing between neighboring ASes. After presenting Abilene and GEANT in Section 3, we introduce our methodology for correlating BGP routing changes that propagate between neighboring ASes in Section 4. Section 5 quantifies the BGP routing changes that propagate between them and their propagation time. We end in Section 6 with a summary of our main findings and a discussion of their implications.

2 BGP in Neighboring ASes

Neighboring ASes connect in one or more physical locations, which we call *interconnection points*. Figure 1 illustrates two neighboring ASes X and Y, where $x1, x2, x3, x4$ and $y1, y2, y3, y4$ are routers in X and Y, respectively, and $p1, p2, p3, p4$ are destination prefixes. X and Y have two interconnection points $(y1, x3)$ and $(y2, x4)$.
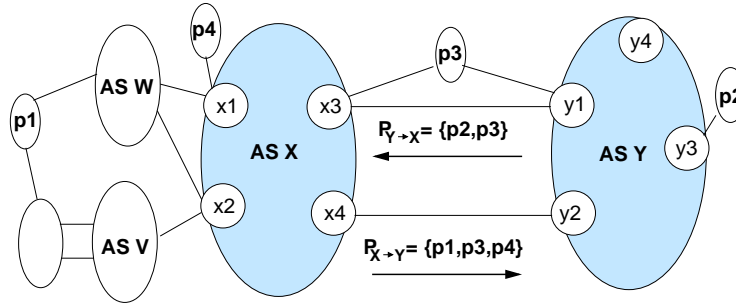

Fig. 1. Route propagation between two neighboring ASes X and Y.

Routers at interconnection points exchange reachability information to destination prefixes using external BGP (eBGP). We use the notation $P_{X \rightarrow Y}$ to refer to the set of prefixes that X announces to Y (even if Y is not using the route learned from X to that prefix). In the example, $P_{X \rightarrow Y} = \{p1, p3, p4\}$, even though Y might use the path to $p3$ it receives from elsewhere, instead of the route from X. BGP routing changes in X for prefixes that belong to $P_{X \rightarrow Y}$ may propagate to Y via the interconnection points.

A router can also learn a BGP route for a destination prefix from other routers in its own AS using internal BGP (iBGP). For example, router $y3$ learns the route to $p3$

from router $y1$. Each router selects the best route to reach this prefix using a multi-step decision process [8], which compares routes based on local policy preferences for path attributes (to a destination) such as AS-path length. Routes learned at all interconnection points to a neighboring AS often have the same AS path length, and other identical attributes. For example, X may learn equally-good routes to $p1$ at $x1$ and $x2$. We call each border router that receives a best route to reach a prefix p from eBGP an *egress router* for p, and the set containing all the egress routers for p as the *egress set* for p. For example, the egress set for $p1$ at AS X is composed of $x1$ and $x2$. Routers inside the AS break the tie among the routes learned from each router in the egress set by selecting the BGP route from the closest (in an intra-domain sense) egress router. This decision is commonly called *hot-potato* routing.

3 Abilene and GEANT

Abilene and GEANT are quite different networks. GEANT is an Internet service provider dedicated to academic institutions, whereas Abilene is a private academic network that is not connected to the commercial Internet. As we will see, these differences allow us to illustrate the impact of engineering decisions and network connectivity on route propagation.

3.1 Inter-Connectivity

Abilene and GEANT have a peer relationship to exchange traffic between their respective customers. Since Abilene is not an Internet provider, all networks that connect to Abilene must have a separate connection to the Internet, by which they can also reach GEANT's customers. GEANT, on the other hand, has six connections to the commercial Internet. GEANT routers have BGP tables with approximately $170,000$ destination prefixes, whereas the BGP tables for Abilene routers have slightly under $10,000$ prefixes.

Research and academic institutions in Europe connect to GEANT through national or regional research networks. Some of these national academic networks have their own connectivity to commercial ISPs. On the other hand, Abilene connects directly to individual institutions. Because of its connection policy, GEANT has many more opportunities for route aggregation, which explains why Abilene announces to GEANT twice as many BGP prefixes than GEANT to Abilene ($|P_{A \to G}| = 5,770$ whereas $|P_{G \to A}| = 2,200$).

Abilene and GEANT have two peering links: between Washington DC (WA) and Frankfurt (DE2), and between New York (NY) and Amsterdam (NL). Abilene and GEANT announce BGP routes with equal AS-path length in both peering locations, and use the same local preference value in both locations as well. Consequently, each router selects between the two interconnection points using hot-potato routing. Neither of the two networks use BGP's Multi-Exit Discriminator (MED) attribute.

3.2 Measurement Infrastructure

Both Abilene and GEANT use Juniper routers running a full-mesh of iBGP sessions. BGP monitors in both networks are NTP synchronized. However, their measurement

infrastructure differs significantly. Abilene has one Zebra BGP monitor per PoP. Given that there is only one router per PoP in Abilene, each monitor establishes an iBGP session as a *client* of the router and collect periodic table dumps as well as all BGP messages reporting changes to the best route to each prefix. The union of BGP messages from all routers gives a global view of each router's choice of best routes for each prefix. GEANT uses a single Zebra BGP monitor that participates in the iBGP full mesh. This monitor is configured as an iBGP peer and thus only receives BGP messages reporting routes learned from eBGP. It does not receive BGP update messages triggered by internal routing changes.

4   Measurement Methodology

This section describes our methodology to correlate BGP update measurements in neighboring ASes. For these correlated BGP changes, we also compute the time it takes until the BGP change in X causes a change in Y, and vice-versa.

4.1   Classification of BGP Changes

First, we classify BGP routing changes from the vantage point of each AS according to the three categories described below, which are inspired from [3]. The main distinction between our work and [3] is that they evaluate the impact of routing changes at X on X's traffic, whereas we classify the BGP routing changes in an AS X that propagate to a neighboring AS Y[4].

Prefix down in X. When X looses connectivity to a prefix $p \in P_{X \to Y}$, each border router sends a message reporting the withdrawal of p. This withdrawal may impact Y in two different ways: routers in Y also withdraw p or shift to another route that does not use X.

Prefix up in X. Similarly, when X gains connectivity to a prefix $p \in P_{X \to Y}$, each border router x sends an announcement of p. Routers in Y may experience a prefix up as well, in the case Y did not have a route to p before receiving the update message; or an egress-set change to use the route from X.

Egress-set change in X. We define an *egress-set change* as a BGP event that changes the composition of the egress set for a given prefix. Routers in X can still reach p, but decide to change routes because the previous route was withdrawn or a new (better) route came up. There are three different types of egress-set changes in X: a change to a worse, equivalent, or better route. For example, suppose that the link between AS W and $p1$ fails in Figure 1. X would then replace this route with the one through AS V, which is worse than the previous route because it has an AS-path length of two, instead of a length of one via W. This change would not trigger an egress-set change in Y, because even though the new route via X is worse, Y does not have a better alternative. In the case that Y has another route to $p1$ that is better than the new route via X, then Y would change routes to $p1$.

The collection of BGP update messages from all routers in an AS contains a lot of redundancy. Indeed, multiple routers report the same routing change, and a single router

---

[4] Although intra-domain routing changes can also impact neighboring ASes because of hot-potato routing [1] or cold-potato routing [8], we do not consider these type of changes here.

may also send multiple messages for the same prefix in a very short period of time because of path exploration [5]. The main classification challenge is therefore to extract one instance of each BGP routing change from all BGP update messages. We extract BGP routing changes using the methodology described in [3]. For each prefix, we group all BGP routing changes that happen close in time [1, 3]. For the results presented in this paper, we select a 70-second threshold to eliminate redundant BGP update messages (approximately 75% of the BGP updates, which is consistent with [3]). We use the timestamp of the first BGP update in the group of updates that leads to a BGP routing change as the timestamp for the change.

4.2 Correlating BGP Routing Changes

Given a time series of labeled BGP routing changes and a time window T, we determine which of the BGP routing changes at Abilene propagate to GEANT, and vice-versa. We call an AS X the *source* of a change, if the routing change happens first at X, and then propagates to Y (which we call the *destination*). We develop a routing correlation algorithm that proceeds in two steps:

Selection of relevant BGP routing changes. We measure $P_{A \to G}$ using BGP table snapshots and BGP messages collected at GEANT. Since we want $P_{A \to G}$ to contain *any* prefix that might be announced by Abilene to GEANT during our analysis, we search for any destination prefix that has at least one BGP message with next-hop AS equal to Abilene's. Similarly, we search Abilene's BGP messages to extract $P_{G \to A}$.

If $P_{A \to G} \bigcap P_{G \to A} \neq \emptyset$, then the causal relationship between BGP routing changes to a destination prefix $p \in P_{A \to G} \bigcap P_{G \to A}$ is not clear. In fact, each AS should use its direct route to p most of the time, except for transient periods of failures. Therefore, we exclude all prefixes in $P_{A \to G} \bigcap P_{G \to A}$ from our analysis to focus on the set of *distinct* destination prefixes that Abilene announces to GEANT, and vice-versa.

Matching related BGP routing changes. Our algorithm first reads the stream of BGP routing changes of Y and creates a list of time-ordered changes per destination prefix. Then, we identify whether each BGP routing change of X triggered a change in Y. For each BGP routing change for a prefix p in X of type c at time t, we search the list of changes to p in increasing time order. We say that a change in X triggered another in Y of type c' at time t', if $t \leq t' \leq t + T(p)$ and c' is *compatible* with c. We define compatibility as follows. Two routes are *compatible* if the type of BGP routing change at the source and destination ASes falls into one of the categories in Table 1. This algorithm returns the list of BGP routing changes of X, where each change is annotated with the corresponding change in Y or a null value.

Type at source AS	Type at destination AS
prefix down	prefix down
	egress-set change
prefix up	prefix up
	egress-set change
egress-set change	egress-set change

Table 1. Compatibility of BGP routing changes at neighboring ASes.

Given the frequent churn of BGP messages caused by events at several locations in the Internet, any heuristic to match BGP routing changes at neighboring ASes has the risk of mistakenly correlating two BGP routing changes that did not propagate between the neighbors in question. Take the example in Figure 1 and suppose that Y uses another neighbor (not shown in the figure) to route to $p1$. A failure at $p1$'s network could cause a prefix down both at X and Y, even though the BGP routing change did not propagate from X to Y. Our algorithm would mistakenly correlate these routing changes. Although we leave a detailed study of these false matches for future work, we include some tests in our algorithm to reduce the likelihood of these false matches:

– **Selection of the prefixes to consider**. We search BGP tables from both networks to determine $P_{A \to G}$ and $P_{G \to A}$ and remove prefixes in the intersection, which could lead to false matches.
– **Classification of BGP routing changes**. We ensure that only compatible routing changes are correlated.
– **Selection of time window** T. The time window guarantees that events that happen too far apart do not get correlated. We set this time window to the worst-case propagation time between the two neighbors. By using the worst-case propagation time, we guarantee to find all truly correlated BGP routing changes while limiting the number of false matches. The next section explains the procedure to find the worst-case propagation time from network configuration data and BGP tables.

4.3 Worst-Case Time Propagation

Our correlation algorithm searches for BGP routing changes that happen *close* in time at both networks, where close means within a time window T. Since routing configurations are different in Abilene and GEANT, we choose a different time window $T_{A,G}$ from Abilene to GEANT and $T_{G,A}$ from GEANT to Abilene. We define the time window as the worst-case BGP propagation time among all the interconnection links.

The propagation of a BGP message is influenced by iBGP (to transfer the message from the egress router to the interconnection point), and eBGP (to transfer the message between the interconnection routers). Juniper routers use an "out-delay" timer to avoid sending updates too often. eBGP sessions may also apply the "route-flap damping" mechanism upon the reception of BGP messages coming from an external neighbor. Another important factor is router load. A measurement study of BGP "pass-through" times [9] showed that the number of prefixes advertised in a BGP session and the number of BGP peers are key contributors to router load. Propagation time also depends on other properties such as network propagation delay and route reflector hierarchy, but neither are relevant here.

The propagation time from GEANT to Abilene has two main components: an out-delay and a load-related delay. GEANT sets the out-delay at the interconnection sessions 10 seconds at (NL,NY) and 30 seconds at (DE2,WA). The worst-case scenario for the transfer delay would be a reset of one of the sessions with GEANT's providers. In this case, the transfer of the $170,000$ routes to one iBGP neighbor should take around 3 minutes [1]. If the router issuing the updates is CPU bound, which is usually the case when it has to treat a large number of updates, then it will send updates to each neighbor sequentially. If the BGP monitor was the last iBGP neighbor to receive the updates,

then it would only receive an update reporting a change after all the other 21 neighbors (i.e., 21×3 minutes after the interconnection point received the change). Therefore, we bound the propagation time of events from GEANT to Abilene with a one-hour time window.

The time window for the propagation of routing events from Abilene to GEANT is mainly determined by route-flap damping imposed by GEANT at the reception of updates from Abilene. GEANT sets the maximum delay introduced by route-flap damping mechanism according to the RIPE recommendations [10] (i.e., 30, 45, and 60 minutes for short, medium, and long prefixes). Abilene does not set the out-delay timers and there is little load-related delay (Abilene's largest BGP session is with GEANT, and it only has $2, 200$ prefixes). Therefore, we use an adaptive time window that depends on the prefix length: $T_{A,G}$ is 1820 seconds, if prefix is shorter than /22; 2720 seconds, if prefix is /22 or /23; and 3620 seconds, if prefix is longer than /24.

5 Analysis of Route Propagation

We now analyze each pair of BGP routing changes from Abilene to GEANT, and vice-versa, correlated according to the methodology described in Section 4. First, we characterize which kinds of BGP routing changes are more frequent and therefore have a more significant impact between Abilene and GEANT. Then, we quantify the route propagation time.

5.1 Classification of Propagated Routes

Table 2 presents the number of *BGP routing changes* per type as defined in Section 4.1. The first half of the table presents the number of BGP routing changes at the source AS. The second half quantifies the *impact* of these changes on the destination AS.

BGP routing change			**Impact**		
Type	Abilene	GEANT	Type	Abilene to GEANT	GEANT to Abilene
prefix down	19, 109	4, 318	prefix down	5, 496	1, 506
			egress-set change	3, 636	94
prefix up	22, 262	6, 214	prefix up	7, 467	2, 558
			egress-set change	4, 803	316
egress-set change	6, 925	3, 591	egress-set change	82	0
total	48, 296	14, 123	total impact	21, 484	4, 474

Table 2. BGP routing changes that propagate between Abilene and GEANT.

Abilene experienced $48, 296$ BGP routing changes that could potentially impact GEANT during the measurement period, whereas GEANT only experienced $14, 123$ BGP routing changes that could impact Abilene. This difference is explained by a combination of two factors: (i) the number of prefixes in $P_{A \to G}$ is more than twice the number in $P_{G \to A}$, and (ii) Abilene does not apply any delay to filter BGP messages, which leads to a higher number of BGP routing changes. Both sets of results show that prefix up and down events dominate the routing changes of each network (these events represent 85.7% of events at Abilene and 74.6% at GEANT).

The first line of Table 2 shows that there were $19,109$ prefix-down events at Abilene that could impact GEANT, but that less than half of those ($9,132$) actually triggered a BGP routing change at GEANT. One reason is that the route-flap damping mechanism applied by GEANT filters many of these events. Another reason is that GEANT can also reach most prefixes that it learns from Abilene using its own connection to the commercial Internet. If GEANT is not using the route via Abilene, then the loss of reachability in Abilene does not impact GEANT.

Given the limited number of alternative paths that Abilene has to reach the prefixes announced by GEANT and vice-versa, most egress-set changes at the source AS have no impact at the destination. In particular, Abilene has almost no alternative to reach the prefixes announced by GEANT ($P_{G \to A}$). Therefore, even when a prefix goes down or if GEANT changes to a worse route, Abilene routers have no alternative but to lose connectivity to the prefix or still select the route they learn from GEANT, respectively. There are only 94 instances in which a prefix down at GEANT caused Abilene to replace its egress set to the prefix. We have verified some of these events manually and observed that Abilene and GEANT have some common peers (mainly research and educational networks in Latin America, Asia-Pacific region, and Africa). Some prefixes are multi-homed to GEANT and to one of these other peers. Abilene uses either one of the routes to reach these prefixes, and events at GEANT cannot impact Abilene when it is using the route via the other peer. This behavior explains why only $1,600$ out of the $4,318$ prefix down at GEANT trigger a change in Abilene.

Table 2 illustrates the types of routing changes that propagate between these two academic peers. We expect the types of routing changes to vary substantially for different pairs of neighboring ASes, because of their relationship, the number of connections to their neighbors, and their location in the Internet hierarchy. GEANT and Abilene experience mostly gain or loss of reachability (or, "prefix up" and "prefix down"). Both networks are fairly small and are close to the edge networks, which implies that they are closer to the network that originates the BGP routing change and that there is less aggregation of prefixes. These results are in sharp contrast with the 6.0% of loss/gain of reachability measured at a tier-1 ISP network [3]. The majority of events at the tier-1 network were distant/transient disruptions, which we classify as egress-set changes.

5.2 Propagation Time

We estimate the propagation time of a routing change to a prefix p between Abilene and GEANT by comparing the time BGP monitors at each network receive the *first* BGP message that reports the routing change to p. We compute the propagation time from a source to a destination AS as the difference between the time of the BGP routing change at the destination and the time at the source. Figure 2 presents the cumulative distribution of the propagation time of all correlated BGP routing changes. (Note that the x-axis is in log scale.)

Although the propagation time in both directions is less than one minute for approximately half of the correlated BGP routing changes, the shapes of the curves are strikingly different. Abilene does not use out delay, therefore over 35% of GEANT BGP routing changes triggered by Abilene happen within the first second in GEANT. The linear increase of the propagation time is an effect of a combination of the route-flap
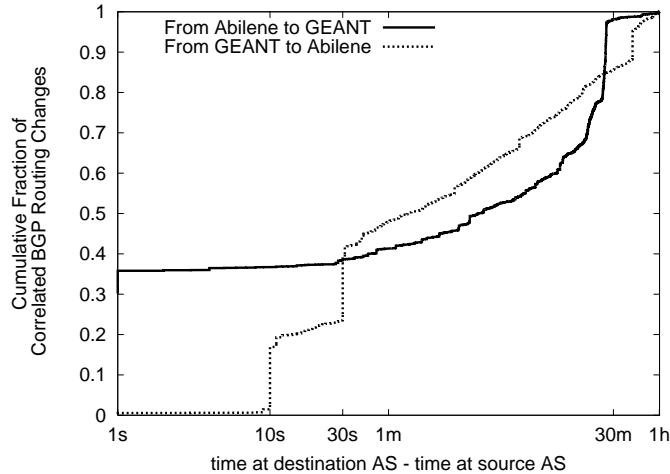
Fig. 2. Propagation time of routing changes.

damping mechanism to enter GEANT, and of other load-related variations (as examined in [9]). The propagation time reaches a plateau at around 30 minutes, which we suspect is due to false matches (there are less than 2% of correlated events with more than 30 minutes propagation time).

On the other hand, the analysis of the propagation time from GEANT to Abilene shows that almost all BGP routing changes from GEANT take at least 10 seconds to reach Abilene. Indeed, the distribution of propagation time has two distinguishable steps at 10 and 30 seconds, which correspond to the out delay of 10 seconds imposed by GEANT at the NL router and of 30 seconds at DE2. The propagation time of almost half of BGP routing changes from GEANT to Abilene is determined by these timers. The slow increasing slope is due to the interaction of a number of factors: TCP behavior at the BGP session, the CPU load at the border router, and the number of BGP messages triggered by each BGP routing change.

We examined the small steps in the distribution from GEANT to Abilene that appear around 3, 7, and 40 minutes propagation times. We found that all of these steps correspond to BGP session resets. A session reset triggers a large number of BGP messages. All these messages reach the neighboring AS at approximately the same time, and consequently have similar propagation times. We conjecture that the few BGP routing changes with propagation time over 30 minutes are due to the load in the GEANT router that first experiences the change (as discussed in Section 4). If the BGP monitor is among the first peers to be notified of a large session reset and the interconnection points to Abilene are among the last ones, we expect time lags even larger than 30 minutes. For instance, the sharp increase in the time propagation distribution around 40 minutes from GEANT to Abilene happens because of the re-establishment of a session with one of the providers. Certainly, 40 minutes of propagation time between neighboring networks is extremely large, but also rare. This example illustrates the importance of taking into account the router load as a factor of propagation time. Events such as

session resets or hot-potato routing changes can trigger thousands of routes to change at the same time [1, 3], and hence substantially increase the load in the router.

6   Conclusion

This paper shows that BGP route propagation is most sensitive to engineering and connectivity of the networks it traverses. The propagation of BGP routing changes between neighboring ASes can sometimes take more than 30 minutes. The longest propagation times from Abilene to GEANT are due to route-flap damping. From GEANT to Abilene, the highest propagation times are caused by the load of the router where routes are processed. GEANT has BGP sessions where it learns more than 150,000 prefixes from a neighboring AS. A reset of any of these sessions would generate a prohibitively large number of BGP updates that would in turn impact the router's load. Note that any AS that has a provider can experience a similar phenomenon, because ASes learn full BGP tables in the session with their providers. The number of prefixes exchanged in each BGP session and the number of BGP sessions per router are important factors that impact router load.

Acknowledgments

References

1. R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of Hot-Potato Routing in IP Networks," in *Proc. ACM SIGMETRICS*, June 2004.
2. S. Agarwal, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Impact of BGP Dynamics on Intra-Domain Traffic," in *Proc. ACM SIGMETRICS*, June 2004.
3. J. Wu, Z. Mao, J. Rexford, and J. Wang, "Finding a needle in a haystack: pinpointing significant BGP routing changes in an IP network," in *Proc. USENIX Symposium on Networked Systems Design and Implementation*, May 2005.
4. D. Pei and J. V. D. Merwe, "BGP convergence in MPLS VPNs," in *Proc. Internet Measurement Conference*, 2006.
5. C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet Routing Convergence," *IEEE/ACM Trans. Networking*, vol. 9, pp. 293–306, June 2001.
6. Z. M. Mao, R. Govindan, G. Varghese, and R. Katz, "Route Flap Damping Exacerbates Internet Routing Convergence," in *Proc. ACM SIGCOMM*, August 2002.
7. R. Oliveira, B. Zhang, D. Pei, R. Izhak-Ratzin, and L. Zhang, "Quantifying path exploration in the internet," in *Proc. Internet Measurement Conference*, 2006.
8. S. Halabi and D. McPherson, *Internet Routing Architectures*. Cisco Press, second ed., 2001.
9. A. Feldman, H. Kong, O. Maennel, and A. Tudor, "Measuring BGP pass-through times," in *Proc. of Passive and Active Measurement Workshop*, pp. 267–277, 2004.
10. C. Panigl, J. Schmitz, P. Smith, and C. Vistoli, "Recommendations for coordinated route-flap damping parameters," October 2001. `http://www.ripe.net/ripe/docs/routeflap-damping.html`.