On the sensitivity of transit ASes to internal failures

Steve Uhlig[*]

Computing Sciences and Engineering department
Université catholique de Louvain, Louvain-la-neuve, Belgium

Abstract. Network robustness is something all providers are striving for without being able to know all the aspects it encompasses. A key aspect of network robustness concerns its sensitivity to internal failures. In this paper, we describe a methodology allowing an AS to study the sensitivity of its network to router and link failures. We provide an open-source tool implementing the sensitivity model of [1], allowing network operators to study the sensitivity of their network to internal failures.
We apply our methodology on the GEANT network and show that some of the routers and links of GEANT are sensitive to internal failures in terms of the control plane. Our results indicate that improvements can be made to the network design so as to reduce the risk of disruptions due to internal failures.

Keywords: network design, sensitivity analysis, BGP, IGP.

1 Introduction

Designing robust networks is a complex problem. Network design consists of multiple, sometimes contradictory objectives. This problem has been fairly discussed in the literature, in particular [2,3]. Examples of desirable objectives during network design are minimizing the latency, dimensioning the links so as to accomodate the traffic demand without creating congestion, adding redundancy so that rerouting is possible in case of link or router failure and, finally, the network must be designed at the minimum cost. In this paper, we focus on single link and node failures because recent papers have shown that large transit networks might be sensitive to internal failures. In [4], Teixeira et al. have shown that a large ISP network might be sensitive to hot-potato disruptions. [5] extended the results of [4] by showing that a large tier-1 network can undergo significant traffic shifts due to changes in the routing. To measure the sensitivity of a network to hot-potato disruptions, [1] has proposed a set of metrics that capture the sensitivity of both the control and the data planes to internal failures inside a network.

To understand why internal failures are critical in a large transit AS, it is necessary to understand how routing in a large AS works. Routing in an Autonomous System (AS) today relies on two different routing protocols. Inside an AS, the intradomain routing protocol (OSPF [6] or ISIS [7]) computes the shortest-path between any pair of routers inside the AS. Between ASes, the interdomain routing protocol (BGP [8]) is used to exchange reachability information. Based on both the BGP routes advertized

[*] Steve Uhlig is "chargé de recherches" with the FNRS (Fonds National de la Recherche Scientifique, Belgium).

by neighboring ASes and the internal shortest paths available to reach an exit point inside the network, BGP computes for each destination prefix the "best route" to reach this prefix. For this, BGP relies on a "decision process" [9] to choose its a single route called the "best route among several available ones. The "best route" can change for two reasons. Either the set of BGP routes available has changed, or the reachability of the next-hop of the route has changed due to a change in the IGP. In the first case, it is either because some routes were withdrawn by BGP itself, or that some BGP peering with a neighbor was lost by the router. In the second case, any change in the internal topology (links, nodes, weights) might trigger a change in the shortest path to reach the next hop of a BGP route. In this paper we consider only the changes that consist of the failure of a single node or link inside the AS, not routing changes related to the reachability of BGP prefixes.

In this paper, we propose a methodology and an open-source tool allowing network operators to study the sensitivity of their network to internal failures. Contrary to [1] whose implementation of the sensitivity model is not available, our tool is freely available and will be integrated in the TOTEM [10] toolbox. We rely on the metrics proposed in [1] and extend the model by removing the limitations on the structure of the BGP sessions inside the AS as well as considering the complete BGP decision process [9]. Furthermore, while [1] studied the sensitivity of the control plane of a tier-1 AS, here we study the sensitivity of the control plane of the GEANT network.

The remainder of this paper is structured as follows. In Section 2 introduces the methodology used to build snapshots of a transit ISP routing and traffic matrix. In Section 3 we present the data collected from the GEANT network. Section 4 introduces the building blocks of the sensitivity model. Section 5 presents the metrics to measure the control plane sensitivity. Section 6 applies these metrics on the control plane of GEANT.

2 Methodology

To build snapshots of real ISP networks and study the impact of internal changes on the changes in the best routes, we rely on CBGP [11], an open-source routing solver. The main point of our relatively heavy methodology is to make the model as easy as possible to match with the context of real transit ASes. We do not make assumptions on the internal graph of the iBGP sessions, even though in the case of GEANT there is a iBGP full mesh between all border routers. The route solver on which we rely, CBGP [11], has no restriction on the structure of the iBGP sessions inside an AS. CBGP has been designed to help the evaluation of changes to the design of the BGP routing inside an AS. Changes to the routing policies of an AS, or the internal configuration of its iBGP sessions is easy with CBGP. Finally, our tool is open-source, and will be integrated in the TOTEM toolbox [10]. Further description of how to use CBGP to model a transit AS can be found in [12].

The most closely related works from the literature are [13] and [14]. The aim of [13] was to provide the networking industry with a software system to support traffic measurement and network modeling. This tool is able to model the intradomain routing and study the implications of local traffic changes, configuration and routing. [13] does not model the interdomain routing protocol though. [14] proposed a BGP emulator that computes the outcome of the BGP route selection process for each router in a single AS. This tool does not model the flow of the BGP routes inside the AS, hence it does not reproduce the route filtering process occurring within an AS. None of these two tools are publicly available. To our knowledge, there is no publicly available implementation of the sensitivity model proposed in [1].

3 Data collected from the GEANT network

GEANT is the pan-European research network and it is operated by Dante. It carries research traffic from the European National Research and Education Networks (NRENs) connecting universities and research institutions. GEANT has POPs (Point of Presence) in all the European countries[1]. All the routers of GEANT are border routers. The backbone of GEANT is composed of 23 POPs and 38 core links interconnecting the POPs.

GEANT captures netflow statistics [15] from all ingress interfaces of its backbone. We used 28 days of data, starting on November 25th 2004 10 AM. GEANT uses a 1/1000 sampling rate to limit the size of the netflow statistics. In the literature, sampling is sometimes done on already sampled netflow, and/or a subset of the traffic is used [16]. To limit the processing burden during the computation the traffic matrices, we processed the raw Netflow and aggregated the flows inside each Netflow file on a source IP - destination IP basis, both IPs having their last 8 bits set to 0. Among all destination prefixes, we computed the largest destination prefixes in traffic volume that accounted for 90% of the total traffic. We relied on these 4911 destination prefixes only in the following simulations.

For this study, we used the routing data available from GEANT. The routing data of GEANT is accessible on request at [17]. The ISIS of GEANT is captured at a collector in Geneva using PyRT [18]. GEANT has chosen to collect all the iBGP routes by having a single Zebra bgpd collector located in Geneva inside the iBGP full mesh with the 23 POPs.

4 Network sensitivity to internal failures

In the remainder of this paper, we propose an implementation of the sensitivity model proposed in [1]. Our method is abstract and generic in that it can be appplied to any network without restrictions on the BGP configuration of the network. We present the model as well as the important aspects related to how we implemented it, and then apply the model on the GEANT network. In this section we introduce the notation and the building blocks necessary for the metrics that capture the sensitivity of changes in the topology of a network on its control plane (Section 5).

Let $G = (V, E, w)$ be a graph, V the set of its vertices, E the set of its edges, w the weights of its edges. A graph transformation δ is a function $\delta : (V, E, w) \rightarrow (V', E', w')$ that deletes vertices or edges from G. In this paper we consider only the

[1] An overview map of the GÉANT network is publicly available from `http://www.geant.net/upload/pdf/Topology_Oct_2004.pdf`.

graph transformation δ that consist in removing a single vertex or edge from the graph. For consistency with [1], we denote the set of graph transformations of some class (router or link failures) by $\varDelta G$. The new graph obtained after applying the graph transformation δ on the graph G is denoted by $\delta(G)$. Due to space limitations, we restricted the set of graph transformations as well as the definition of a graph compared to [1], as we do not study the impact of changes in the IGP cost. Changes to the IGP cost occur rarely in real networks, and never in the GEANT network. Our methodology however has no limitation on the set of graph transformations, IGP changes could be considered simply by extending our definition of a graph G and adding the corresponding set of graph tranformations.

To perform the sensitivity analysis to graph transformations, one must first find out for each router how graph transformations may impact the egress point it uses towards some destination prefix p. The set of considered prefixes is denoted by P. The BGP decision process $dp(v, p)$ is a function that takes as input the BGP routes known by router v to reach prefix p, and returns the egress point corresponding to the best BGP route. The *region index set* RIS of a vertex v records this egress point of the best route for each ingress router v and destination prefix p, given the state of the graph G: $RIS(G, v, p) = dp(v, p)$.

We introduced the state of the graph G in the *region index set* to capture the fact that changing the graph might change the best routes of the routers. The next step towards a sensitivity model is to compute for each graph transformation δ (link or router deletion), whether a router v will shift its egress point towards destination prefix p. For each graph transformation δ, we recompute the all pairs shortest path between all routers after having applied δ, and record for each router v whether it has changed its best BGP route towards prefix p. We denote the new graph after the graph transformation δ as $\delta(G)$. As BGP advertisements are made on a per-prefix basis, the best route for each (v, p) pair has to be recomputed for each graph transformation. It is the purpose of the *region shift function* H to record the changes in the egress point corresponding to the best BGP route of any (v, p) pair, after a graph transformation δ:

$$H(G, v, p, \delta) = \begin{cases} 1, \ if \ RIS(G, v, p) \neq RIS(\delta(G), v, p) \\ 0, \ otherwise \end{cases}$$

The *region shift function* H is the building block for the metrics that will capture the sensitivity of the network to the graph transformations.

To summarize how sensitive a router might be to a set of graph transformations, the *node sensitivity* η computes the average *region shift function* over all graph transformations of a given class (link or node failures), for each individual prefix p:

$$\eta(G, \varDelta G, v, p) = \sum_{\delta \in \varDelta G} H(G, v, p, \delta) \cdot Pr(\delta)$$

where $Pr(\delta)$ denotes the probability of the graph transformation δ. Note that we assume that all graph transformations within a class (router or link failures) are equally likely, i.e. $Pr(\delta) = \frac{1}{|\varDelta G|}, \forall \delta \in \varDelta G$, which is reasonable unless one provides a model for link and node failures. Further summarization can be done by averaging the *vertex sensitivity* over all vertices of the graph, for each class of graph transformation. This gives the *average vertex sensitivity* $\hat{\eta}$:

$$\hat{\eta}(G, \Delta G, p) = \frac{1}{|V|} \sum_{v \in V} \eta(G, \Delta G, v, p)$$

The *node sensitivity* is a router-centric concept that performs an average over all possible graph transformations. Another viewpoint is to look at each individual graph transformation δ and measure how it impacts all routers of the graph on average. The *impact of a graph transformation* θ is computed as the average over vertices of the *region shift function*:

$$\theta(G, p, \delta) = \frac{1}{|V|} \sum_{v \in V} H(G, v, p, \delta)$$

The *average impact* of a graph transformation $\hat{\theta}$ summarizes the information provided by the *impact* by averaging it over all graph transformations of a given class:

$$\hat{\theta}(G, \Delta G, p) = \sum_{\delta \in \Delta G} \theta(G, p, \delta) \cdot Pr(\delta)$$

5 Control plane sensitivity

Section 4 provided the basic notions to deal with the sensitivity of the network to graph transformations. In this section, we present the metrics of [1] that measure the impact of graph transformations on the control plane.

[1] relied on a worst-case sensitivity and a best-case one in their *region shift function*, to capture the uncertainty as to whether a graph transformation would lead to a change of the egress point of a route for sure or not, dependingon the behavior of the actual tie-breaking rules of the BGP decision process. In this paper, the *region shift function* relies on the BGP decision process as it exists on most routers [9], corresponding to a situation in-between the worst-case and best-case ones used in [1]. All the metrics defined in this section will have *RM* in superscript to indicate that these metrics concern the *routing matrix*, i.e. the set of egress points that can be used to reach a destination prefix by each ingress router.

In practice, the same egress points can be used for several destination prefixes by a router. The impact of a graph transformation δ can thus affect many destination prefixes. To capture the impact of a graph transformation on the number of prefixes that will have to change their egress point, we sum for each graph transformation, the values of the *region shift function* over all considered prefixes and divide it by the total number of prefixes:

$$H^{RM}(G, P, v, \delta) = \frac{1}{|P|} \sum_{p \in P} H(G, v, p, \delta)$$

This new function H^{RM} is called the *routing shift function* for the control plane.

Based on the *routing shift function* for the control plane, we can now define the routing sensitivity of routers to graph transformations: the *node routing sensitivity*. The *node routing sensitivity* η^{RM} is computed as for each router, the sum of the values of the *routing shift function* (for the control plane) over all values of the graph transformations multiplied by the graph transformation probabilities:

$$\eta^{RM}(G, P, \mathit{\Delta} G, v) = \sum_{\delta \in \mathit{\Delta} G} H^{RM}(G, P, v, \delta) \cdot Pr(\delta)$$

Again, we consider that all graph transformations are equally likely so that $Pr(\delta) = \frac{1}{|\mathit{\Delta} G|}$. The *average node routing sensitivity* $\hat{\theta}^{RM}$ summarizes the node routing sensitivity by doing the average of the *node routing sensitivity* over all routers:

$$\hat{\eta}^{RM}(G, P, \mathit{\Delta} G) = \frac{1}{|V|} \sum_{v \in V} \eta^{RM}(G, P, \mathit{\Delta} G, v)$$

While the *node routing sensitivity* η^{RM} provides an average over all graph transformations, a desirable goal for network design is to try to minimize the impact of the routing shifts at any router. To know the worst graph transformation in terms of the routing shift at each node, we compute the *worst routing shift* η^{RM}_{max} for each node, i.e. the maximum of the *routing shift function* over all graph transformations:

$$\eta^{RM}_{max}(G, P, \mathit{\Delta} G, v) = \max_{\delta \in \mathit{\Delta} G} H^{RM}(G, P, v, \delta)$$

For network robustness, one does not only care about the impact of the graph transformations on any single router of the network, but also the impact of a specific node or router failure on the whole network. For this, the *routing impact of a graph transformation* θ^{RM} is computed as the average fraction of route shifts (H^{RM}) over all vertices:

$$\theta^{RM}(G, P, \delta) = \frac{1}{|V|} \sum_{v \in V} H^{RM}(G, P, v, \delta)$$

The *average routing impact* $\hat{\theta}^{RM}$ summarizes the *routing impact* by averaging its value over the set of graph transformations of each class:

$$\hat{\theta}^{RM}(G, P, \mathit{\Delta} G) = \sum_{\delta \in \mathit{\Delta} G} \theta^{RM}(G, P, \delta) \cdot Pr(\delta)$$

Network design is not only about trying to minimize the average impact of link and node failures, but also the impact of the worst failure inside the network. The *maximum routing impact of a graph transformation* θ^{RM}_{max} gives for each graph transformation, the largest value of H^{RM} over all possible vertices of the graph:

$$\theta^{RM}_{max}(G, P, \delta) = \max_{v \in V} H^{RM}(G, P, v, \delta)$$

Finally, the *network-wide worst routing impact* σ^{RM}_{max} gives the routing sensitivity for the router of the graph most impacted by any graph transformation of each class:

$$\begin{aligned} \sigma^{RM}_{max}(G, P, \delta) &= \max_{v \in V} \eta^{RM}_{max}(G, P, \mathit{\Delta} G, v) \\ &= \max_{\delta \in \mathit{\Delta} G} \theta^{RM}_{max}(G, P, \delta) \end{aligned}$$

It is not always possible to prevent parts of the network from being vulnerable to large routing shifts, without incurring huge costs to provide redundancy. Knowing the most vulnerable parts of the network allows network operators to rely on techniques like protection or fast rerouting [19] to ensure the availability of routing paths under specific failures.

6 Control plane sensitivity of the GEANT network

In this section we apply the metrics defined in the previous section on the control plane of the GEANT network. Figure 1 presents the *routing impact of the graph transformations* (θ^{RM}) on the routers of the GEANT network. The top part of Figure 1 gives the impact of router failures while the bottom part gives the impact of (bidirectional) link failures. Our study relies on 28 daily snapshots of the life of GEANT, so each error bar on the graphs of Figure 1 gives the min-average-max (indicated by a point, beginning of continuous line, end of continuous line) values over the 28 time bins of the study. For all figures that display on their x-axis either routers or graph transformations, the objects shown represented in the x-axis have been ordered by increasing values of their average impact or sensitivity over time. The y-axis of Figure 1 gives the *routing impact* in percentage of the considered prefixes that shift their egress point after the failure.

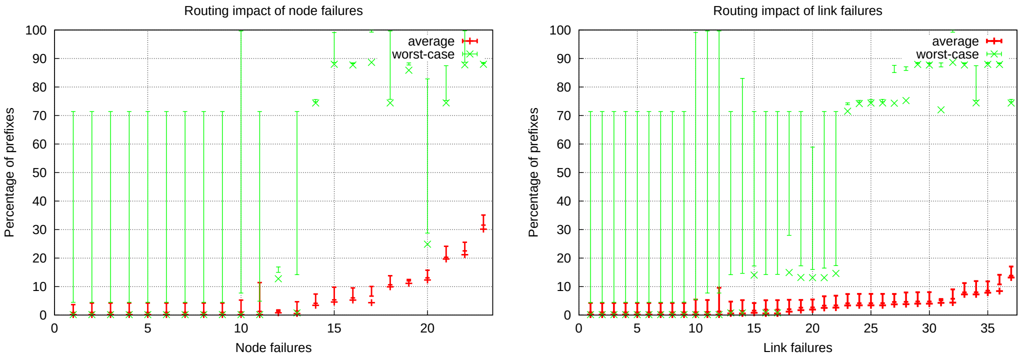

Fig. 1. Routing impact to graph transformations (θ^{RM}): router (left) and link (right) failures.

Let us start with the *routing impact* of node failures (top part of Figure 1). The average *routing impact* of node failures is very small, under 5%, for most of them. The worst node failure (θ^{RM}_{max}) impacts on average about 30% of the routes. To have a small average impact for a graph transformation means that the concerned routers or links are not used very often as egress points by the routers of the network. We can see that only 6 routers seem critical in the GEANT topology in that respect. In the GEANT network, some routers are mainly used to connect the NRENs to the network, not to provide connectivity outside the NRENs. These routers only attached to NRENs and not other peers are mainly ingress points and are not used much as egress points by other routers of the network. Their failure hence mostly impacts the connectivity with a few prefixes advertised by the concerned NREN. On the other hand, some routers can have a non-negligible routing impact in the network. The *worst-case routing impact* (θ^{RM}_{max}) is more complex than the average routing impact. The graph transformations having a small routing impact also have a small *worst-case routing impact* most of the time, except for one particular time bin (valid for router and link failures). The graph transformations that have the largest routing impact however have a large *worst-case routing impact* all the time, meaning that these graph transformations are critical for at least one router all the time. This means that the router or link concerned by these critical graph transformations will be highly disruptive for at least one router of the network. Improving the resilience of the network could hence be done by protecting these routers that might suffer from these highly disruptive graph transformations, or by splitting the best routes of these routers so reduce the impact of a single router or link failure.

The observations made so far relate to the design of the GEANT network which relies a lot on hot-potato routing and where no BGP tweaking is made so as to split the set of best routes used to reach prefixes evenly among the available egress points of the network.

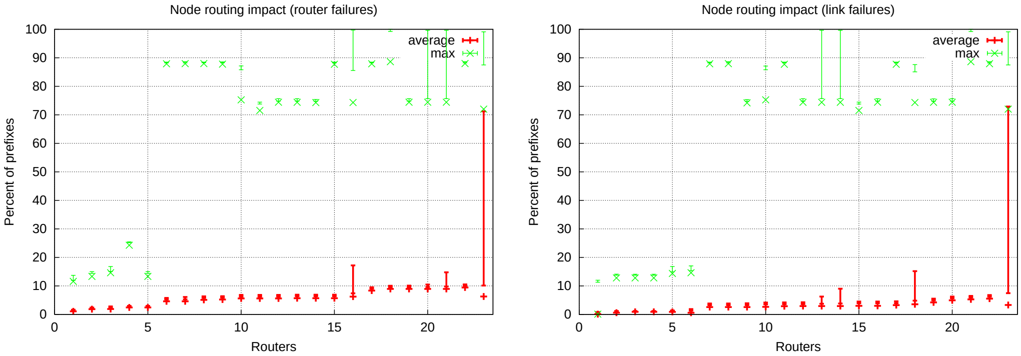

Fig. 2. Node sensitivity to graph transformations (η^{RM}): router (left) and link (right) failures.

While the *routing impact* gives an average over the routers of the network, it is interesting to have a more detailed view at the individual sensitivity of each router of the topology to graph transformations, with the *node routing sensitivity* (η^{RM}). Figure 2 shows for each router of the GEANT network, the *node routing sensitivity* (η^{RM}) for each router, along with the *worst routing shift* (η^{RM}_{max}). Figure 2 shows that the average sensitivity is small, and more evenly balanced among the routers that the impact of the graph transformations on Figure 1. Only one router suffered from a large average *routing impact*, but only for a single time bin. So if we assume that all graph transformations are equally likely, the risk that a given router will suffer from big routing shifts is low on average. However, the *worst routing shift* (η^{RM}_{max}) tells us another story. All except a few routers will suffer a very large routing shift (more than 70% of its routes) for at least one graph transformation, meaning that all the best routes of that router cross the concerned link or router. This does not forcibly mean that the network design is bad, but that improvement in the design could be made by trying to spread the best routes over the available paths and egress points of the network to prevent a single link or router failure to have such a large impact on some routers.

Even though some graph transformations are more important than others (particularly router failures) when their impact is averaged over all routers, individual routers do not see wide differences in their average sensitivity to graph transformations. The situation for the *worst-case routing impact* (θ^{RM}_{max}) and the *worst-case node routing sensitivity* (η^{RM}_{max}) is quite different. Almost all routers on Figure 2 show a large *worst-case node routing sensitivity* (η^{RM}_{max}), meaning that most routers are highly impacted by at least one graph transformation, even though on average each router is not much affected by graph transformations. This point to the fact that with BGP, large set of prefixes share the same egress point for a given ingress router. Hence it is highly likely that at least one router or link failure will affect an important egress point for any given router. Note that a few routers are not very sensitive to graph transformations. These nodes are actually those having external peerings, i.e. the routers most heavily used as egress points in the network. As these routers very often have as their best route one learned from an external peer, they are those most insensitive to disruptions that occur inside the network. The five routers that are the less sensitive to link and router failures are actually those that are most critical for all the rest of the network. This means there is room for improving the design of the network by reducing the ciriticality of these five routers, at least by splitting the best routes of the ingress routers more evenly between these five egress routers so that one failure does not impact so much some routers.

7 Conclusions and further work

In this paper we extended the sensitivity model proposed in [1] to understand to what extent it allows to understand the robustness of a transit AS. We proposed a methodology to make this study reproducible on other large ISP networks. Then, we described how we implemented our version of the sensitivity model. Our tool can help ISPs to design their network and improve its robustness. Our version of the model is sensitive to any predicted change of the best BGP route selected by a router, and does not rely on assumptions concerning the internal BGP configuration of the network.

We applied the sensitivity analysis on the GEANT to better understand its design and robustness. Our analysis showed that some of the routers and links of the GEANT network are highly critical and sensitive to internal failures. This has implications on the protection that might be done inside the network to prevent critical router and link failures to create big disruptions in the network. Furthermore, we found great consistency between the results of the control plane and the data plane (not shown due to lack of space), indicating that applying the analysis on the control plane might be sufficient to provide insight into the design of the network. As collecting traffic information is a very demanding task [16], especially for large transit networks, ISPs might benefit from our methodology by doing the same analysis as carried in [1] and this paper, solely based on the routing information that is much easier to collect and analyze. We already applied this analysis on a large tier-1 AS network, which provided a more critical view of the insight given by the sensitivity analysis [20].

8 Acknowledgments

We thank GEANT for making their data available. Thanks to Renata Teixeira for her insightful comment on earlier draft versions of this work. This work was partially supported by the E-NEXT NoE funded by the European Commission. This research was carried while the author was visiting Intel research Cambridge.

References

1. R. Teixeira, T. Griffin, G. Voelker, and A. Shaikh, "Network sensitivity to hot potato disruptions," in *Proc. of ACM SIGCOMM*, August 2004.
2. R. S. Cahn, *Wide Area Network Design: Concepts and Tools for Optimisation,* Morgan Kaufmann, 1998.
3. W. D. Grover, *Mesh-Based Survivable Networks,* Prentice Hall PTR, 2004.
4. R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of hot-potato routing in IP networks," in *Proc. of ACM SIGMETRICS*, June 2004.
5. R. Teixeira, N. Duffield, J. Rexford, and M. Roughan, "Traffic matrix reloaded: impact of routing changes," in *Proc. of PAM 2005*, March 2005.
6. J. Moy, *OSPF : anatomy of an Internet routing protocol*, Addison-Wesley, 1998.
7. D. Oran, "OSI IS-IS intra-domain routing protocol," Request for Comments 1142, Internet Engineering Task Force, Feb. 1990.
8. J. Stewart, *BGP4 : interdomain routing in the Internet*, Addison Wesley, 1999.
9. Cisco, "BGP best path selection algorithm," `http://www.cisco.com/warp/public/459/25.shtml`.
10. "TOTEM: a TOolbox for Traffic Engineering Methods," `http://totem.info.ucl.ac.be/`.
11. B. Quoitin, "C-BGP, an efficient BGP simulator," `http://cbgp.info.ucl.ac.be/`, September 2003.
12. B. Quoitin and S. Uhlig, "Modeling the routing of an Autonomous System with CBGP," *To appear in IEEE Network Magazine, special issue on interdomain routing*, 2005.
13. Anja Feldmann, Albert Greenberg, Carsten Lund, Nick R eingold, and Jennifer Rexford, "NetScope: Traffic Engineering for IP Networks," *IEEE Network Magazine*, March 2000.
14. N. Feamster, J. Winick, and J. Rexford, "A model of BGP routing for network engineering," in *Proc. of ACM SIGMETRICS*, June 2004.
15. Cisco, "NetFlow services and applications," White paper, available from `http://www.cisco.com/warp/public/732/netflow`, 1999.
16. G. Varghese and C. Estan, "The measurement manifesto," *Comput. Commun. Rev.*, vol. 34, no. 1, pp. 9–14, 2004.
17. "Intel-DANTE monitoring project," `http://www.cambridge.intel-research.net/monitoring/dante/`.
18. R. Mortier, "The Python Routeing Toolkit (PyRT)," `http://ipmon.sprint.com/pyrt/`, 2001.
19. J.-P. Vasseur, M. Pickavet, and P. Demeester, *Network Recovery: Protection and Restoration of Optical, SONET-SDH, and MPLS*, Morgan Kaufmann, 2004.
20. S. Uhlig and S. Tandel, "A critical view of the sensitivity of ases to internal failures," in *Proc. of the 2005 Tyrrhenian International Workshop on Digital Communications, Sorrento, Italy*, July 2005.