On the Sensitivity of Transit ASes to Internal Failures

Steve Uhlig*

Department of Computing Science and Engineering
Université catholique de Louvain, Louvain-la-neuve, B-1348, Belgium
suh@info.ucl.ac.be

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 design is the sensitivity of the network to internal failures. In this paper we present 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 we show that some of the routers and links of GEANT are sensitive to internal failures. Our results indicate that improvements can be made to the network design so as to reduce the risk of disruptions due to internal failures. Furthermore, we show great consistency between the results of the control plane and the data plane, indicating that applying the analysis on the control plane might be sufficient to provide insight into how to improve the resilience of the network to internal failures.

Keywords: network design, sensitivity analysis, control and data planes, BGP, IGP.

1 Introduction

Designing robust networks is a complex problem. Network design consists of multiple, sometimes contradictory objectives [2, 3]. Examples of desirable objectives during network design are minimizing the latency, dimensioning the links so as to accommodate 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. Recent papers have shown that large transit networks might be sensitive to internal failures. [4] has 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

* This research was carried while the author was visiting Intel research Cambridge. Steve Uhlig is “chargé de recherches" with the FNRS (Fonds National de la Recherche Scientifique, Belgium)

T. Magedanz, E.R.M. Madeira, and P. Dini (Eds.): IPOM 2005, LNCS 3751, pp. 142–151, 2005.

used to exchange reachability information. Based on both the BGP routes advertised 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 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 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. 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 both the control and the data planes of the GEANT network.

Our study confirms that the metrics proposed in [1] provide insight into the sensitivity of the network to internal failures. More important is the necessity to confront the sensitivity analysis of the control and the data planes of [1], because the uneven traffic distribution towards destination prefixes [10, 11] might make the results of the control and data planes different. Our study of the GEANT network however indicates that the control and the data plane of this network have a similar sensitivity.

Note that for reasons of space limitation we do not describe in this paper the methodology used to build snapshots of the routing and traffic of an AS, but we refer the reader to [12].

The remainder of this paper is structured as follows. Section 2 introduces the building blocks of the sensitivity model. Section 3 presents the metrics to measure the control plane sensitivity. Section 4 applies these metrics to the control plane of GEANT. Section 5 then presents the metrics to measure the data plane sensitivity and Section 6 studies the sensitivity of the data plane of GEANT.

2 Network Sensitivity to Internal Failures

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 vertex or edge from G. In this paper we consider only graph transformations 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 ΔG. 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], by not considering 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 with weights and adding the corresponding set of graph transformations.

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:

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

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. $Pr(\delta) = \frac{1}{|\Delta G|}, \forall \delta \in \Delta G$, 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:

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

3 Control Plane Sensitivity

[1] relied on worst-case and best-case sensitivities 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, depending on 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.

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:

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

Again, we consider that all graph transformations are equally likely so that $Pr(\delta) = \frac{1}{|\Delta G|}$. 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:

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

While the *node routing sensitivity* 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_{max}^{RM} for each node, i.e. the maximum of the *routing shift function* over all graph transformations:

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

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:

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

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_{max}^{RM} gives for each graph transformation, the largest value of H^{RM} over all possible vertices of the graph:

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

4 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. For this study, we used the largest prefixes that account for 90% of the total traffic of GEANT during the 28 considered days, a total of 4911 prefixes.

Figure 1 presents the *routing impact of the graph transformations* (θ^{RM}) on the routers of the GEANT network. The left part of Figure 1 gives the impact of router failures while the right one gives the impact of link failures. Our study relies on 28 daily snapshots in 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 days 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.

The left part of Figure 1 provides the *routing impact* of node failures. The average *routing impact* of node failures is very small, under 5%, for most of them. The worst

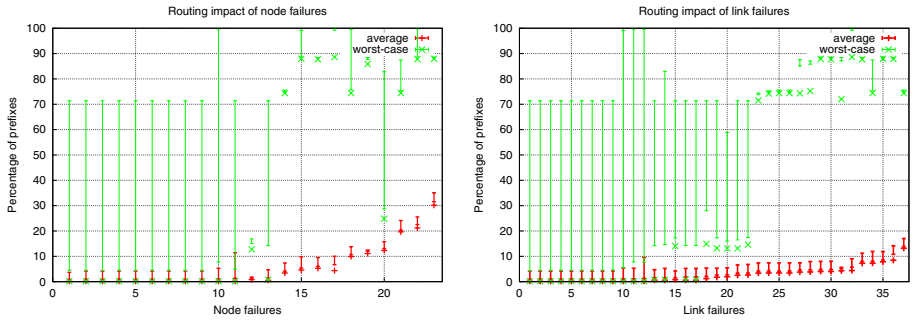

Fig. 1. Routing impact to graph transformations: router (left) and link (right) failures

node failure (θ_{max}^{RM}) impacts on average about 30% of the 4911 considered prefixes. 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 (National and Regional research and Education Networks) 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* is a little more complex to understand 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. 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.

Note that the observations made so far are highly related 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.

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. Figure 2 shows for each router its *node routing sensitivity* (η^{RM}), along with the *worst routing shift* (η_{max}^{RM}). 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_{max}^{RM}) gives us another

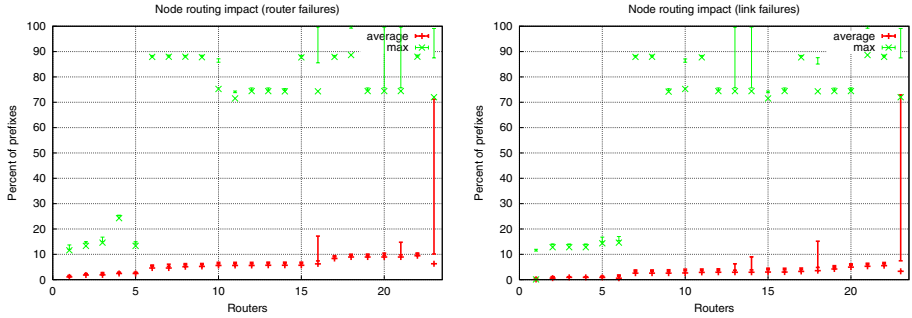

Fig. 2. Node sensitivity to graph transformations: router (left) and link (right) failures

viewpoint. 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 implies that improvement in the design can 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_{max}^{RM}) and the *worst-case node routing sensitivity* (η_{max}^{RM}) is quite different. Almost all routers on Figure 2 show a large *worst-case node routing sensitivity*, 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 less sensitive 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 criticality 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.

5 Data Plane Sensitivity

Let P be the set of destination prefixes and $I \in V$ be the set of ingress routers. The traffic demand M is an $|I|$ x $|P|$ matrix, whose elements $M(v,p)$ represent the amount of traffic that is received at ingress router v towards destination prefix p. The total inbound traffic received at an ingress router towards all destination prefixes of P is

$$T(v) = \sum_{p \in P} M(v,p)$$

In this paper we use one-day time bins for the traffic demand. We do not index all variables by the time to prevent unnecessarily cumbersome notations, but the reader must be aware that all variables are computed for each time bin. Similarly to the previous section, all metrics of this section have TM in superscript to indicate that they concern the traffic matrix.

As the *routing shift function* H^{RM} for the control plane, the *traffic shift function* H^{TM} gives for each prefix p, the amount of traffic entering ingress v that switches to other egress routers after a graph transformation δ. This is done by summing over all prefixes $p \in P$, the value of the *region shift function* H multiplied by the amount of traffic for the given (v, p) pair:

$$H^{TM}(G, P, v, \delta) = \frac{1}{T(v)} \sum_{p \in P} H(G, v, p, \delta) \cdot M(v, p)$$

The sensitivity of each ingress router to traffic shifts is represented by the *ingress node traffic sensitivity* η^{TM} and is computed as the sum over all graph transformations of the *traffic shift function* H^{TM} multiplied by the probability of the graph transformation δ:

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

Each transformation is again supposed to be equally likely. The *maximal ingress node traffic sensitivity* η^{TM}_{max} is, for each ingress node, the maximum of the *traffic shift function* over all possible graph transformations:

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

Then the *average ingress node traffic sensitivity* $\hat{\eta}^{TM}$ gives the average of the *ingress node traffic sensitivity* computed over all ingresses, for each graph transformation:

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

The *traffic impact of a graph transformation* θ^{TM} measures the fraction of the traffic that shifts because of a graph transformation δ, averaged over all ingress points of the graph:

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

θ^{TM} captures the change in the traffic matrix due to the graph transformation. The *maximal traffic impact* θ^{TM}_{max} of a graph transformation δ gives the maximum of the *traffic shift function* H^{TM} computed over the ingress nodes of the graph:

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

The *average traffic impact* $\hat{\theta}^{TM}_{max}$ sums the *traffic impact of a graph transformation* θ^{TM} over all graph transformations δ multiplied by the probability of the graph transformation:

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

6 Data Plane Sensitivity of the GEANT Network

In this section, we go to the data plane side of the sensitivity analysis. As traffic in general seems to be unevenly distributed among the destination prefixes [10, 11], one should not expect that the sensitivity analysis for the control plane be consistent with the one of the data plane.

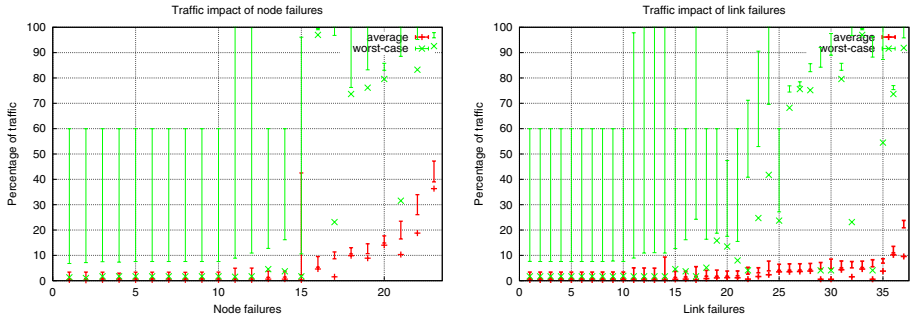

Fig. 3. Traffic impact to graph transformations: router (left) and link (right) failures

Figure 3 shows the *traffic impact of graph transformations*. As usual, the graph transformations on the x-axis of Figure 3 have been ordered by increasing value of their average *traffic impact* over the 28 daily snapshots. The impact of the graph transformations are similar for the data and the control planes. The average impact of the graph transformations are small for most graph transformations. We can see that the most disruptive router failure has a slightly larger average traffic impact than its routing one, about 39% of the traffic against 31.5% of the considered prefixes. But overall, the results for the traffic and routing impact are pretty much the same. The *worst-case traffic impact* (θ_{max}^{TM}), as for the control plane, is smaller than 10% for 14 routers and 21 links except for a single time interval. The consistency between the results for the control plane and the data plane indicate that the distribution of the traffic among ingress-egress pairs inside GEANT samples relatively well the distribution of the egress points found by BGP. The traffic matrix does not seem to change much the routing sensitivity in the GEANT network, at least for the largest 4911 prefixes capturing 90% of the traffic during the 28 days we considered.

7 Conclusions

In this paper we proposed an implementation of the sensitivity model to internal failures of [1]. 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 GEANT to better understand its design and robustness to internal failures. We showed that some of the routers and links of the GEANT network are highly critical and sensitive to internal failures. This analysis 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 consistency between the results of the control plane and the data plane, indicating that applying the analysis on the control plane might be sufficient to provide insight into the design of the network. We believe that large ISPs might benefit from carrying the same study as we did in this paper to improve their understanding of their network design choices.

Acknowledgments

We thank GEANT for making their routing and traffic data available. Thanks to Renata Teixeira for comments on earlier versions of this paper. This work was partially supported by the E-NEXT NoE funded by the European Commission.

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. J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP Routing Stability of Popular Destinations," in *Proc. of ACM SIGCOMM Internet Measurement Workshop*, November 2002.
11. S. Uhlig, V. Magnin, O. Bonaventure, C. Rapier, and L. Deri, "Implications of the Topological Properties of Internet Traffic on Traffic Engineering," in *Proc. of ACM SAC'04*, March 2004.
12. B. Quoitin and S. Uhlig, "Modeling the routing of an Autonomous System with C-BGP," *To appear in IEEE Network Magazine, special issue on interdomain routing*, 2005.