Triangle Inequality and Routing Policy Violations in the Internet

Cristian Lumezanu, Randy Baden, Neil Spring, and Bobby Bhattacharjee

University of Maryland
{lume,randofu,nspring,bobby}@cs.umd.edu

Abstract. Triangle inequality violations (TIVs) are the effect of packets between two nodes being routed on the longer direct path between them when a shorter detour path through an intermediary is available. TIVs are a natural, widespread and persistent consequence of Internet routing policies. By exposing opportunities to improve the delay between two nodes, TIVs can help myriad applications that seek to minimize end-to-end latency. However, sending traffic along the detour paths revealed by TIVs may influence Internet routing negatively. In this paper we study the interaction between triangle inequality violations and policy routing in the Internet. We use measured and predicted AS paths between Internet nodes to show that 25% of the detour paths exposed by TIVs are in fact available to BGP but are simply deemed "less efficient". We also compare the AS paths of detours and direct paths and find that detours use AS edges that are rarely followed by default Internet paths, while avoiding others that BGP seems to prefer. Our study is important both for understanding the various interactions that occur at the routing layer as well as their effects on applications that seek to use TIVs to minimize latency.

1 Introduction

End-to-end latencies in the Internet demonstrate triangle inequality violations (TIV). Evidence from various real world latency data sets shows that more than 5% of the triples and more than half of the pairs of nodes are part of TIVs [1, 2, 3]. TIVs are not measurement artifacts, but a natural and persistent consequence of Internet routing [4].

Triangle inequality violations expose opportunities to improve network routing by offering lower-latency one-hop *detour* [5] paths between nodes. Latency-sensitive peer-to-peer applications, such as distributed online games [6] or VOIP [7], could potentially improve their performance by exploiting TIVs [8, 9]. Consider the TIV in Figure 1, where A, B and C are all peers in the same overlay. Node A could reduce its latency to C by simply routing all traffic addressed to C through B.

Of course, all overlays violate routing policies [10]. Sending traffic along the detour paths, exposed by TIVs, instead of default paths, chosen by BGP, has the potential to disrupt traffic engineering and policy routing in the Internet. In the example above, the path ABC may violate the transit agreements between the ISPs of A, B and C (maybe because B's ISP is a customer of both A's and C's ISPs). Do all shorter detour paths violate policies? Or are they simply not selected by BGP because of its lack of mechanisms to minimize delay? Do detour paths traverse a different set of ASes that makes

S.B. Moon et al. (Eds.): PAM 2009, LNCS 5448, pp. 45–54, 2009.

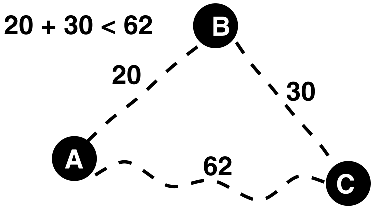


Fig. 1. Example of triangle inequality violation. All latencies are derived from real measurements.

them more attractive to users, but less attractive to ISPs? Answering such questions is important for understanding the effects of exploiting TIVs for end-to-end latency reduction.

In this paper we study the interaction between triangle inequality violations and Internet routing policies. We collect a new, large, real-world latency data set and augment it with measured and predicted AS paths. To the best of our knowledge, this is the first large (1715 nodes) latency data set that contains AS paths between the majority of the nodes. We show that, as one might expect, many of the paths of shorter detours exposed by TIVs appear impossible due to policy routing, but that 25% of them are available to BGP. Our result offers new insight into the effects of latency-reducing overlay routing as well as on how ISPs and end-users can work together to avoid less-than-optimal paths.

Our contributions can be summarized as follows:

- we present a new study on the relationship between triangle inequality violations and routing in the Internet;
- we collect a large symmetric latency data set (1715 nodes) augmented with measured and predicted AS paths; this is the first symmetric data set that contains both measured RTTs and AS paths for all pairs, all collected during the same period of time;
- we show that 25% of the shorter detour paths exposed by TIVs are in fact available to BGP and could potentially provide end-to-end latency reduction without necessarily violating inter-domain policies

The rest of the paper is organized as follows. We discuss related work in Section 2. In Section 3, we present the data collection and methodology. We explore the origins of TIVs in Section 4 and discuss the relationship with BGP in Section 5. We conclude in Section 6.

2 Related Work

Previous research related to triangle inequality violations in the Internet can be grouped into two categories: studies on end-to-end latency [4, 11] and studies on the performance of network coordinate systems [1, 12, 13].

Savage *et al.* [11] measure a large number of Internet paths between geographically diverse hosts and show that alternate paths of lower latency exist between more than

20% of the pairs of nodes in their data sets. The authors study the origins of the TIVs and conclude that the availability of alternate paths does not depend on a few good or bad ASes. We confirm that no individual ASes can influence the latency of a path, but also demonstrate that the way they peer and interconnect with each other can.

Zheng *et al.* [4] use data collected between nodes in the GREN research network to argue that TIVs are not measurement artifacts, but a persistent, widespread and natural consequence of Internet routing policies. We confirm their findings that TIVs are caused by routing policies. We also study and quantify, using much larger latency and AS path data sets, the different policy decisions that may affect the formation of TIVs.

Several studies examine TIVs in relation to the impact they have on network coordinate [2, 14] and positioning [15] systems. Because these systems treat the Internet as a metric space—where TIVs are prohibited—they may obtain inaccurate results. None of these studies [1, 3, 12] considers the interaction between TIVs and Internet routing policies. Understanding the origin and the properties of TIVs would potentially help network coordinate systems to better counter the negative effects of TIVs.

3 Methodology

In this section, we describe the methodology for collecting our latency data set as well as for determining the AS paths between the nodes.

3.1 Latency Data Set

We use King [16] to compute RTTs between 1715 hosts in the Gnutella network. King uses recursive DNS queries to estimate the propagation delay between two hosts as the delay between their authoritative name servers. The IP addresses of the 1715 nodes in our measurement are provided by the Vivaldi project [2]. They were chosen such that the IPs share the same subnet with their authoritative name servers so that better-connected DNS servers would not influence the latency estimates. We run King for all pairs of IPs from a computer at University of Maryland for a week in March 2008. For each pair of nodes we keep the median of all measured latencies.

3.2 AS Paths

Understanding the AS paths beneath the TIV allows us to evaluate the detour routes for their preferences toward "better" ASes or inter-AS connections, or their compliance with known interdomain policies, *i.e.*, whether enhanced BGP protocols might find these detours and thus eliminate the TIVs. To compute as many AS paths as possible between the pairs of nodes in our latency data set we use several sources: RouteViews, Looking Glass servers and *iPlane* [17]. To the best of our knowledge this is the first large latency data set between Internet hosts augmented with AS path information computed at the same time.

RouteViews [18] collects and archives BGP routing tables and updates from commercial ISPs. We gathered AS path information from 44 BGP core routers located in 38 ISPs in March 2008. In addition, we used paths obtained by Madhyastha *et al.* [17] by probing around 25,000 BGP prefixes from 180 public Looking Glass servers.

We augment RouteViews and Looking Glass measured paths with paths predicted by *iPlane*. *iPlane* measures paths from 300 PlanetLab sites to more than 140,000 BGP prefixes to predict end-to-end paths between any pair of hosts. The predicted path combines partial segments of known paths, exploiting the observation that routes from nearby sources tend to be similar [19].

We found AS paths for the pairs of nodes in the data set, 10.4% from RouteViews and 13.6% from Looking Glass. The reason for such low completeness is that most of the Looking Glass servers and RouteViews peers are close to the core of the Internet and are unlikely to capture paths between two edge ASes. *iPlane* predicts AS paths between 71.7% of the pairs. By combining RouteViews, Looking Glass, and *iPlane*, we find AS paths for almost 75% of the pairs of nodes in the data set.

4 Origins of TIVs

To better understand TIVs and their interaction with policy routing, we must gain more insight into how they come into existence. We refer to the default path between two nodes as the direct path (or the long side of the triangle) and to the shorter path through an intermediary as the detour path (or the short sides of the triangle).

First, we look at the AS edge distribution of both direct and detour paths and show that TIVs appear not because poor ASes or AS edges are avoided by detour paths, but because detour paths are able to find better AS edges than the default direct paths. Whether these edges are known to BGP or not we discuss in the next section.

Second, we show that most latency reduction on detour paths is obtained by relaying through nodes that are either close to the source or the destination. By deviating slightly from the default path [20], one can avoid congested peering points or override routing policies that may have inflated the default path in the first place [21].

4.1 AS to AS Edge Usage

We hypothesize that triangle inequality violations appear because packets traversing the detour paths find somehow better AS edges while avoiding overloaded or circuitous edges present on the direct paths. Savage *et al.* proposed a similar hypothesis [11] on the usage of ASes on detour paths. After studying latencies between 39 traceroute servers located in North America, they showed that, for most ASes, the difference between the number of direct and detour paths in which they appeared was low. They concluded that the availability of alternate detour paths does not depend on a few good or poor ASes. We suggest that, although no individual ASes can influence the latency of a path, the way they peer and interconnect with each other can.

To study how the edges traversed by detour and direct paths are preferred or avoided, we define the weight w of an AS edge e:

$$w(e) = \frac{R(e) - D(e)}{R(e) + D(e)}, \forall \text{ edge } e$$

where $R(e)$ is the number of times e is traversed by a detour path and $D(e)$ is the number of times it is traversed by a direct path. w takes values between -1, meaning that

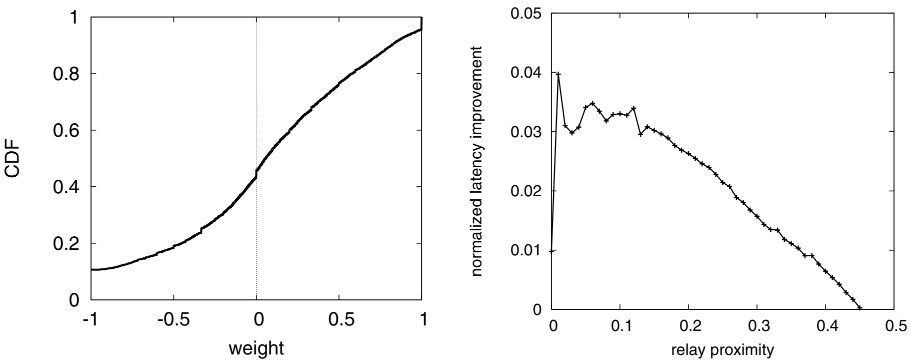


Fig. 2. (left) Distribution of weight for AS-to-AS edges: Positive values correspond to edges that are used more by detours; negative values correspond to edges that are used more by direct paths. **(right)** Distribution of normalized latencies based on the proximity of the relay to source/destination.

detours never use the edge, and 1, meaning that *direct paths* never use the edge. More generally, negative weights indicate that detours avoid the edge and positive weights indicate that detours prefer the edge.

For each pair of nodes in our data set, we find all triangle inequality violations, then select at random one *significant TIV*: a bad triangle where the detour path reduces latency over the direct path by at least 10 ms and 10%. The random selection of one bad triangle per node pair intends to not bias the results toward pathological senders or receivers, or toward the properties of the most severe violations, which might differ from the rest. We compute the weight for every edge that appears at least in one significant TIV and plot the cumulative distribution in Figure 2(left). The vertical line represents a hypothetical situation where each AS edge would be equally used by detours and direct paths. The distribution of weight is, instead, much less balanced. Detours use about 40% of all edges twice as often than the direct path does. About 10% of edges are avoided: they appear in direct paths but are never used by detours.

Thus, using preferred AS edges rather than avoiding the others is key to TIVs. One might expect that detour routing is predominantly about avoiding pathological AS paths. Our result suggests that this intuition is false: using the preferred edges is what gives detour paths lower latency.

4.2 Relay Proximity

For all TIVs in our data set, we analyze the relationship between the location of the intermediate nodes (relays) and the severity of the violation. Intuitively, if the most severe violations are obtained when relay nodes are close to either the source or the destination, then TIVs are likely due to path diversity at the endpoints.

We define the *relay proximity* of a detour path as the ratio between the latency from the relay to the closest endpoint (either the source or the destination) and the latency of the direct path associated to the detour. The relay proximity has values between 0 and 0.5; a value of 0.5 means that the relay is located at half the distance between the endpoints of the path. For each pair of nodes in our data sets, we select the detour corresponding to a significant TIV and compute its relay proximity. We group each detour according to its relay proximity and compute the total latency reduction obtained by each group. For ease of presentation, we normalize the latency reduction by dividing it to the total number of detours. We plot the results in Figure 2(right).

Most latency reduction comes when the relay is close to one of the end points, so it is likely that detours take advantage of path diversity near the endpoints.

5 Triangle Inequality Violations and BGP

It is not surprising that the BGP path selection process may prefer longer, policy-compliant paths to shorter, policy-violating detour paths. In this section we ask, to what extent are detour AS paths available to BGP?

We separate all AS detour paths in our data sets into two categories: *impossible* and *possible*. A path is *impossible* when it could not have been advertised by a neighbor, possibly because it could not have been advertised by a neighbor's neighbor and so on. Common inter-domain routing rules [22] state that customers should not advertise routes learned from a provider to peers or other providers. This prevents the customer from being used as transit between two of its providers (customer transit). Similarly, routes learned from peers are advertised to customers and not to providers or other peers, preventing peer transit. Otherwise, a path appears *possible*, though traffic engineering or other rules may have led to the selection of an alternate.

To assess whether detour paths traverse possible or impossible paths, we use the AS relationships inferred by CAIDA [23]. Directed AS edges belong to one of four categories: customer-to-provider, provider-to-customer, peer-to-peer and sibling-to-sibling. A policy compliant AS path should have zero or more customer-to-provider edges followed by zero or one peer-to-peer edges, followed by zero or more provider-to-customer edges. Sibling-to-sibling edges may appear anywhere on the path.

Table 1 classifies the detour paths. The row labeled "Unknown" corresponds to the AS paths for which we cannot give an indisputable classification using the AS relationship data set. 58% of the detour paths in the data set are non-compliant (*i.e.*, include customer or peer transit). This is not surprising, since detour paths go through end hosts, which are generally customers and may be in stub ASes. To validate our results, we performed the same classification using only the detour and direct AS paths derived from RouteViews. While the percentage of possible paths decreased only slightly (21%), we obtained more non-compliant paths (70%) and fewer unknown paths (9%). We describe the cells of Table 1 in the following discussion, first for impossible, and then for possible paths.

5.1 How Impossible Are the Impossible Paths?

We ask the following question: How severe are the policy violations of the impossible paths? For each detour path we define its prefix and its suffix. The prefix is the longest common subpath to appear at the beginning of both the detour path and a policy compliant path between the same pair of nodes, while the suffix is the longest common

Table 1. Detour paths are *possible* (may be available to the BGP decision process) or *impossible* (not advertised by BGP). Percentages inside the tables are relative to the total possible or impossible paths. Categories separated by horizontal lines overlap.

Total Detours		793,693
Impossible AS Paths		460,830 (58%)
Cause	Customer transit	343,381 (75%)
	Peer transit	117,449 (25%)
Type	Truly disjoint	302,207 (66%)
	Borderline	153,057 (33%)
	Undercover	5,503 (1%)
Possible AS Paths		197,453 (25%)
Traffic Eng.	Relay AS not on direct path	56,813 (29%)
	Direct, detour paths differ	103,215 (52%)
	Direct, detour paths same	37,425 (19%)
Path length	Shorter than direct	17,770 (9%)
	Equal to direct	75,032 (38%)
	Longer than direct	104,651 (53%)
Transit cost	Smaller than direct	35,541 (18%)
	Equal to direct	96,751 (49%)
	Greater than direct	65,161 (33%)
Unknown		135,410 (17%)

subpath to appear at their end. Based on the prefix and the suffix, we define two measures to capture the severity of policy violation of a detour path: *width* and *depth*. The width is the number of valid AS edges that would be required to connect the suffix and the prefix to obtain a policy compliant path. The depth is the minimum number of AS edges that have to be traversed from the relay to the end of the prefix or the beginning of the suffix. Based on the values of width and depth we classify the impossible detour paths into *undercover*, *borderline*, and *truly disjoint*. We present an example of each type in Figure 3 and describe them below:

undercover (depth = 0) (1% of impossible detour AS paths)

Because the depth is 0, the relay of the detour lies on a compliant path. Although both direct and detour traffic enter the AS of the relay, they use different peering points to exit.

borderline (depth = 1, width $\leq$ 1) (33%)

Borderline compliant detours diverge from the compliant path only to traverse the relay before returning quickly.

truly disjoint (all other cases) (66%)

A truly disjoint path differs from any compliant path by at least two AS edges.

The results above show that one third of the "impossible" paths are within *one AS hop* of being "possible". Enhanced BGP protocols—where nodes exchange path

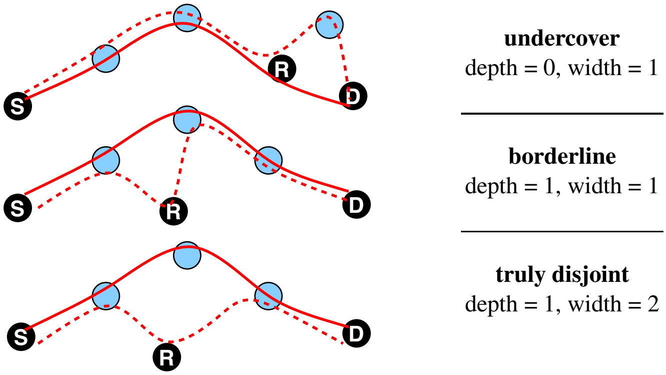


Fig. 3. Examples of impossible detour AS paths

performance information with their neighbors—could learn about these detours, eliminate the TIVs and offer faster paths to end users.

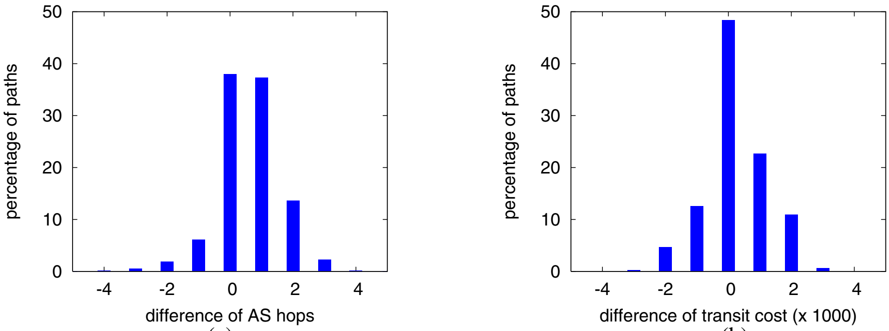
5.2 Possible Paths

25% of the detours in the data set follow compliant paths. Therefore, they can be learned by BGP. Only traffic engineering decisions or a lack of configuration can stop these paths from being advertised and learned. BGP routers select paths based on cost, performance, length, and even which path is advertised first. Since we do not know precisely why any path was chosen, we consider here a few possible explanations.

Traffic Engineering. Each AS must pay some cost to carry traffic in its internal network. ISPs engineer their networks and routing to minimize this cost, while improving performance, choosing early-exit routes that deliver packets at the nearest exit, or divert traffic to balance load. Although we do not have explicit information about these choices, we can infer when such traffic engineering occurs. For example, for 52% of the possible detour paths, the AS of the relay node lies *on the direct path*, yet the detour and the direct paths are different. This may occur because traffic, when redirected through the relay, will traverse a different peering point than the default traffic.

These results suggest that detours may take advantage of shorter paths by overriding common traffic engineering practice. The number of detours due to minimizing internal cost may be higher than we have observed; we can only identify such detours when the relay is on the direct path.

Path Length. When choosing among otherwise equal paths, BGP selects the one with the fewest ASes. Because a detour path traverses an additional relay point, we expect it to use more ASes than the corresponding direct path. For each pair of nodes, we compute the difference in number of AS hops between the detour path and the corresponding direct path. Over 90% of the *possible* detour paths traverse at least as many ASes as the corresponding direct paths (Figure 4(a) and Table 1). This suggests that latency is not reduced by eliminating ASes traversed.



(a) (b)

Fig. 4. Possible detour AS paths have larger (a) path length, and (b) transit cost

Transit Cost. Although not visible in BGP data, the price an ISP pays to its provider may make a path more or less preferred. Traversing larger networks implies greater expense. We define the transit cost of a path as the maximum degree—number of AS-to-AS peerings—of all ASes on the path. Table 1 and Figure 4(b) show the results of the comparison between the transit cost of detour paths and corresponding direct paths. The transit cost of the detour paths is significantly higher than that of direct paths.

6 Conclusions

In this paper, we offer new evidence into both the origins and properties of Internet triangle inequality violations. We show that triangle inequality violations occur because many AS edges are constantly avoided by BGP. By analyzing the decisions that lead to such occurrences, we show that, not surprisingly, most detour paths of TIVs violate interdomain routing policies. ISPs control Internet routing using BGP which chooses paths primarily based on cost, policies, past performance, even which route arrives first, but never based on end-to-end latency. However, we find that 25% of the paths in our data sets are available to BGP (and 20% more are borderline available): BGP knows of low latency paths but prefers them less.

Our intention is not to reprimand BGP for not being able to offer low-latency paths to its users. We show that, in fact, there is room for improvement in Internet routing, without affecting the equilibrium of the tussle [24]: both end-users and ISPs could take advantage of better paths without any of them feeling cheated. We intend to explore in future work ways in which latency-reducing overlay networks and policy routing can coexist.

References

1. Wang, G., Zhang, B., Ng, T.S.E.: Towards network triangle inequality violation aware distributed systems. In: IMC (2007)
2. Dabek, F., Cox, R., Kaashoek, F., Morris, R.: Vivaldi: a decentralized network coordinate system. In: SIGCOMM (2004)

3. Lumezanu, C., Levin, D., Spring, N.: PeerWise discovery and negotiation of faster paths. In: HotNets (2007)
4. Zheng, H., Lua, E.K., Pias, M., Griffin, T.G.: Internet routing policies and round-trip times. In: Passive and Active Measurement Workshop (2005)
5. Savage, S., Anderson, T., Aggarwal, A., Becker, D., Cardwell, N., Collins, A., Hoffman, E., Snell, J., Vahdat, A., Voelker, G., Zahorjan, J.: Detour: A case for informed Internet routing and transport. IEEE Micro. 19(1), 50–59 (1999)
6. Bharambe, A., Douceur, J.R., Lorch, J.R., Moscibroda, T., Pang, J., Seshan, S., Zhuang, X.: Donnybrook: Enabling large-scale, high-speed, peer-to-peer games. In: ACM SIGCOMM (2008)
7. Kho, W., Baset, S.A., Schulzrinne, H.: Skype relay calls: Measurements and experiments. In: IEEE Global Internet Symposium (2008)
8. Lumezanu, C., Baden, R., Levin, D., Spring, N., Bhattacharjee, B.: Symbiotic relationships in Internet routing overlays. In: NSDI (2009)
9. Andersen, D.G., Balakrishnan, H., Kaashoek, M.F., Morris, R.: Resilient overlay networks. In: SOSP (2001)
10. Qiu, L., Yang, Y.R., Zhang, Y., Shenker, S.: On selfish routing in Internet-like environments. In: ACM SIGCOMM (2003)
11. Savage, S., Collins, A., Hoffman, E., Snell, J., Anderson, T.: The end-to-end effects of Internet path selection. In: SIGCOMM (1999)
12. Lee, S., Zhang, Z.L., Sahu, S., Saha, D.: On suitability of euclidean embedding of internet hosts. In: Sigmetrics (2006)
13. Lua, E.K., Griffin, T., Pias, M., Zheng, H., Crowcroft, J.: On the accuracy of the embeddings for Internet coordinate systems. In: IMC (2005)
14. Ng, T.S.E., Zhang, H.: Predicting Internet network distance with coordinates-based approaches. In: INFOCOM (2002)
15. Wong, B., Slivkins, A., Sirer, E.G.: Meridian: A lightweight network location service without virtual coordinates. In: SIGCOMM (2005)
16. Gummadi, K., Saroiu, S., Gribble, S.: King: Estimating latency between arbitrary Internet end hosts. In: IMW (2002)
17. Madhyastha, H.V., Isdal, T., Piatek, M., Dixon, C., Anderson, T., Krishnamurthy, A., Venkataramani, A.: iPlane: An information plane for distributed services. In: USENIX OSDI (2006)
18. RouteViews: Routeviews (2008), `http://www.routeviews.org`
19. Madhyastha, H.V., Anderson, T., Krishnamurthy, A., Spring, N., Venkataramani, A.: A structural approach to latency prediction. In: IMC (2006)
20. Yang, X., Wetherall, D.: Source selectable path diversity via routing deflections. In: ACM SIGCOMM (2006)
21. Spring, N., Mahajan, R., Anderson, T.: Quantifying the causes of path inflation. In: ACM SIGCOMM (2002)
22. Gao, L.: On inferring autonomous system relationships in the Internet. IEEE/ACM Transactions on Networking 9(6), 733–745 (2001)
23. Dimitropoulos, X., Krioukov, D., Fomenkov, M., Huffaker, B., Hyun, Y., kc claffy, R.G.: As relationships: inference and validation. SIGCOMM CCR 37(1), 29–40 (2007)
24. Clark, D.D., Wroclawski, J., Sollins, K.R., Braden, R.: Tussles in cyberspace: Defining tomorrow's Internet. IEEE/ACM Transactions on Networking 13(3), 462–475 (2005)