Internet Routing Policies and Round-Trip-Times

Han Zheng, Eng Keong Lua, Marcelo Pias, and Timothy G. Griffin

University of Cambridge Computer Laboratory
{han.zheng, eng.lua, marcelo.pias, timothy.griffin}@cl.cam.ac.uk

Abstract. Round trip times (RTTs) play an important role in Internet measurements. In this paper, we explore some of the ways in which routing policies impact RTTs. In particular, we investigate how routing policies for both intra- and inter-domain routing can naturally give rise to violations of the triangle inequality with respect to RTTs. Triangle Inequality Violations (TIVs) might be exploited by overlay routing if an end-to-end forwarding path can be stitched together with paths routed at layer 3. However, TIVs pose a problem for Internet Coordinate Systems that attempt to associate Internet hosts with points in Euclidean space so that RTTs between hosts are accurately captured by distances between their associated points. Three points having RTTs that violate the triangle inequality cannot be embedded into Euclidean space without some level of inaccuracy. We argue that TIVs should not be treated as measurement artifacts, but rather as natural features of the Internet's structure. In addition to explaining routing policies that give rise to TIVs, we present illustrating examples from the current Internet.

1 Motivation

Since round trip times (RTTs) play an important role in Internet measurements, it is important to have a good understanding of the underlying mechanisms that give rise to observed values. Measured RTTs are the result of many factors — "physical wire" distance, traffic load, link layer technologies, and so on. In this paper, we explore a class of factors that are often ignored — the ways in which routing policies can impact minimum RTTs.

In particular, we investigate how routing policies for both intra- and inter-domain routing can naturally give rise to violations of the triangle inequality with respect to RTTs. The existence of Triangle Inequality Violations (TIVs) impact two areas of current research, one positively, and the other negatively. For overlay routing [1], TIVs represent an opportunity that might be exploited if the layer 3 routed path can be replaced with one of lower latency using a sequence of routed paths that are somehow stitched together in the overlay. On the other hand, TIVs pose a problem for any Internet Coordinate System (ICS) [2, 3, 4, 5, 6, 7, 8] that attempts to associate Internet hosts with points in Euclidean space so that RTTs between hosts are accurately captured by distances between their associated points. The problem is simply that any three points having RTTs that violate the triangle inequality cannot be embedded into Euclidean space without some level of inaccuracy, since their distances in Euclidean space must obey this inequality. We feel that current work on Internet Coordinates too often treats TIVs as measurement artifacts, either ignoring them entirely or arguing that they are not important. We have come to

C. Dovrolis (Ed.): PAM 2005, LNCS 3431, pp. 236–250, 2005.

the opposite conclusion — we feel that TIVs are natural and persistent features of the Internet's "RTT geometry" and must somehow be accommodated. We illustrate how TIVs can arise from routing policies and present illustrating examples from research networks in the Internet. Our measurement results are consistent with those reported in PAM 2004 [9], and indicate that the commercial Internet is even more likely to exhibit such policy-induced TIVs.

2 A Bit of Notation

A *metric space* is a pair $M = (X,\ d)$ where X is a set equipped with the distance function $d : X \to \mathbb{R}^+$. For each $a,\ b \in X$ the *distance between a and b* is $d(a,\ b)$, which satisfies the properties, for all $a,\ b, c \in X$,

(anti-reflexivity) $d(a,\ b) = 0$ if and only if $a = b$,
(symmetry) $d(a,\ b) = d(b,\ a)$,
(triangle inequality) $d(a,\ b) \leq d(a,\ c) + d(c,\ b)$.

A *quasi-metric space* $(X,\ d)$ satisfies the first two requirements of a metric space, but the triangle inequality is not required to hold. This paper argues that Internet RTTs naturally form a quasi-metric space, with routing policies being an important, but not sole, factor in the violation of the triangle inequality.

A Triangle Inequality Violation (TIV) is simply a triple $(a,\ b,\ c)$ that violates the triangle inequality. It is not hard to see that for any TIV, there must be one edge that is longer than the sum of the other two edges.

Suppose that $M_1 = (X_1,\ d_1)$ is a quasi-metric space, and $M_2 = (X_2,\ d_2)$ is a metric space. Every one-to-one function ϕ from X_1 to X_2 naturally defines a metric space called *the embedding of M_1 in M_2 under ϕ*, defined as $\phi(M1) = (\phi(X_1),\ d_2)$. We normally abuse terminology and simply say that ϕ embeds X_1 into X_2, We will be interested in the case where X_1 is a finite set, X_2 is $\mathbb{R}^n$ with the standard notion of Euclidean distance, $d_2(x,\ y) = \|x - y\| = \sqrt{\sum_{1 \leq i \leq n} (x_i - y_i)^2}$.

The number of possible embeddings is quite large. In addition, the *accuracy* of an embedding can be measured in various ways, as is outlined in [10]. In this paper, we will not focus on any particular embedding, nor on any particular notion of accuracy. We simply note that any TIV embedded into Euclidean space must involve some "distortion" since the triangle inequality will hold on the embedded points. Although one might attempt to embed RTT distances into a non-Euclidean space (such as [11]), by far the most common techniques are Euclidean.

3 Routed Paths Versus Round Trip Time

Most data paths are determined by dynamic routing protocols that automatically update forwarding tables to reflect changes in the network. Dynamic routing never happens without some kind of manual *configuration*, and we will refer to routing protocol configuration as implementing *routing policy*. The Internet routing architecture is generally described as having two levels [12] — Interior Gateway Protocols (IGPs) are designed

to route within an autonomously administered routing domain, while Exterior Gateway Protocols (EGPs) route between such domains. In this section we explore the ways in which routing policies can give rise to data paths that violate the triangle inequality with respect to delay.

Intra-domain routing is typically based on finding shortest paths with respect to configured link weights. The protocols normally used to implement shortest path routing are RIP, OSPF, or IS-IS. We note that Cisco's EIGRP [13] presents a slightly more complex routing model, and in addition some networks actually use BGP for intra-domain routing. Nevertheless, for simplicity we will investigate here only how shortest path routing can give rise to TIVs.

In order for there to be no TIVs in shortest path routing, the link weights must be consistent with the actual link delays. However, delay is just one of the many competing demands in the design of intra-domain routing. So the disagreement between the link weight assignment and the actual link delay will cause structural TIVs in the intra-domain case.

We now consider how inter-domain routing can introduce triangle inequality violations (TIVs).

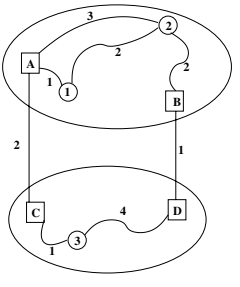

Fig. 1. Nodes 1, 2 and 3 form a TIV due to *Hot Potato Routing*. The numbers on the edges represent link propagation delay

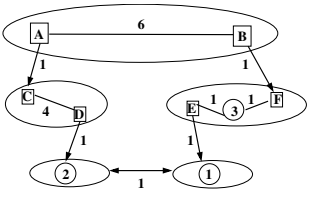

Fig. 2. Nodes 1 and 2 share a "private peering shortcut" that cannot be used to transit traffic between nodes 2 and 3

Hot potato routing [14] refers to the common practice of sending inter-domain traffic towards the closest egress point. Consider two ASes presented as large ovals in Fig. 1. We inspect the "triangle" formed by nodes 1, 2, and 3. The upper AS has two egress points, A and B, the first of which is closer to node 1, while the second is closer to node 2. Node 3 is in the lower AS and is closer to egress C. Note that hot potato routing will result in asymmetric routing between nodes 2 and 3. Traffic from 3 to 1 and 2 will always exit the lower AS at egress point C, whereas traffic from 2 to 3 will exit the upper AS at egress point B. The distance matrix for the nodes 1, 2, and 3, all calculated as "round trip" distance, is

d	1	2	3
1	0	4	8
2	4	0	13
3	8	13	0

Here we see that $13 = d(2,3) > d(2,1) + d(1,3) = 12$, and so this represents a TIV.

Lest the reader think that the problem is asymmetric routing alone, we now show how economic relations between networks can give rise to TIVs even when routing is symmetric. Private peering links are common routing shortcuts used to connect ISPs. Fig. 2 presents five ASes, the upper AS representing a large transit provider, the middle two ASes representing smaller providers, and the lower two ASes representing customers. The directed arrows represent customer-provider relationships pointing from a provider to a customer. The bi-directional arrow between the lower ASes represents a *private peering* [15] link. This link transits only traffic between the lower two ASes, and is not visible to the providers above. (This type of peering is very common on the Internet.) Traffic between nodes 1 and 2 uses this link for a round trip path cost of 2. Traffic between nodes 3 and 1 goes through border router E for a round trip path cost of 4. However, traffic between nodes 2 and 3 must go up and down the provider hierarchy for a round trip path cost of 28!. Here we see that $28 = d(2,3) > d(2,1) + d(1,3) = 6$, so this represents a TIV.

TIV can also be caused by traffic flowing through three independent AS paths between a triple of nodes, where at least some AS along one path is not in the other two paths. This usually happens due to multi-homing [16] and because peering relationship is a bilateral agreement and typically not transitive [17]. In this case, there is absolutely no reason to believe that triangle inequality must hold. We will see some examples of this type of TIV in section 4.

In fact, the interaction between inter-domain and intra-domain routing can also introduce TIVs. This type of TIV applies to the majority of systems that use end-to-end measurement results. This interaction often makes it very difficult to classify the root cause of an observed TIV.

The current inter-domain routing protocol, BGP, conveys only AS-level paths information. Nothing is learned about the actual router-level path within an AS. Therefore, when BGP makes a decision based on the shortest AS path, nothing can be inferred about the actual router-level path. An example of this is shown in Fig. 3.

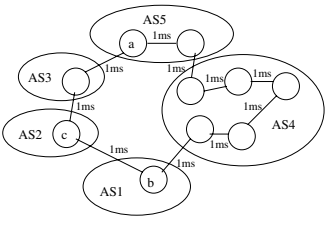

Fig. 3. When choosing the AS-level path between nodes a and b, BGP prefers AS 4 1 to AS 3 2 1, although the router-level path along AS4 is much longer

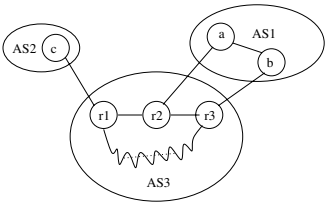

Fig. 4. A TIV caused by the interaction between hot-potato routing and intra-domain TIV

More complicated interactions are also possible. Fig. 4 shows an example of a TIV caused by both hot-potato routing and intra-domain TIV. Routers $r1$, $r2$ and $r3$ form an intra-domain TIV, and AS1 uses hot-potato routing between egress points a and $r2$,

and b and $r3$. So the path between b and c ($b\ r3\ r1\ c$) exhibits a much longer RTT than the paths between a and b, and a and c ($a\ r2\ r1\ c$). We can see that sometimes the intra-domain behavior of the intermediate AS may change the existence of TIV through interaction with inter-domain routing.

4 Case Study: The Global Research and Education Network (GREN)

We define GREN (c.f. [9]) to be all the Autonomous Systems reachable from the Abilene network (AS11537), because we can reach almost all the research and education networks in the world from Abilene, and Abilene has no direct upstream commercial provider [18]. Fig. 5 illustrates the connectivity of most component networks of GREN. We study GREN instead of the commodity Internet because GREN is a relatively more open and transparent network, and we can understand its global structure more easily. In addition, a large percentage of PlanetLab nodes are hosted in GREN networks.

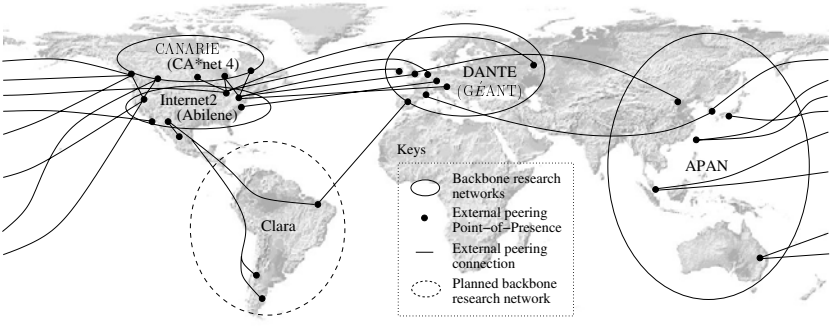

Fig. 5. The connectivity of most component networks of GREN

We take a BGP table dump (on 19 July 2004 at 12:00 GMT) from Oregon Internet Exchange (OIX), a RouteViews server which directly peers with Abilene, and study how GREN is inter-connected. We observe that quite a few commercial ASes are involved to glue together bits of GREN. According to the BGP *behavior* of each AS, we were able to classify all the 1203 GREN ASes into 30 commercial ASes and 1173 research ASes (details omitted). The particular reasons for 'leaking' these commercial ASes into GREN are still under further investigation, although we have seen that some are leaked in for very legitimate reasons.

4.1 Examples of Internal TIVs

We obtained the router-level topology [19] and IS-IS weights from a monitor box inside the GEANT backbone (the multi-gigabit pan-European research network managed by DANTE), as shown in Fig. 6. We also measured the minimum RTT values between all pairs of GEANT backbone routers using their looking glass interface [20]. The RTT

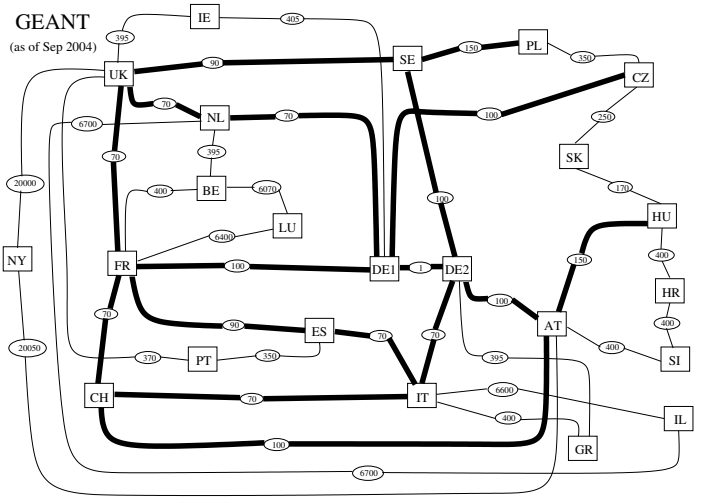

Fig. 6. The topology of GEANT backbone. Thick lines represent OC-192 10Gbps fast backbone links, and other lines represent slower links of various speeds. The numbers in the circles represent the IS-IS weights assigned to the links

measurements are taken 100 times for each pair (50 times starting from each end of the pair), and the minimum is used in our calculations. Our goal in this experimental work is to study the *structural* causes of TIV, therefore we take the minimum measurement to avoid biases in the results due to high variations in RTTs. The measurements are spread out into an 8 hour period, both to smooth out the variations in RTT caused by network conditions, and for rate limiting purposes. The experiment is repeated three times a day for a week from 12 August 2004 to 18 August 2004, so we had 21 RTT matrices. There are 23 backbone routers in the GEANT AS, so each matrix is 23x23 in size. We then obtained the final RTT matrix by taking the minimum measurement of each pair. Out of all the 1771 distinct triangles formed by triples of backbone routers, we observed 244 TIVs . This represents a significant 13.8% TIV inside the GEANT network.

When examined closely, it is observed that the TIVs in the GEANT network are mainly caused by the link weights disproportional to the link delay. For example (see Fig. 7), Slovakia has two OC-48 links to Czech Republic and Hungary, respectively. But their purpose is just to provide access for the Slovakia SANET, not to transit traffic between Czech and Hungary. So the weights on these two access links are intentionally set quite high, so that the traffic from Czech to Hungary would go via an alternative path through Germany and Austria, where the links are backbone OC-192 links with lower weights. When we look at the RTTs between the three nodes, however, the RTT between Czech and Hungary is much larger than the sum of the other two RTTs, causing a TIV.

We then looked at the Abilene network. Abilene publishes their router configuration files online [21], so we obtained their router-level topology and IS-IS weights from their website, as shown in Fig. 8. To verify that their published configuration file is up-to-date, we ran traceroute between directly connected nodes to see that every configured link is actually operational. The configuration data matches very well with the verification. We

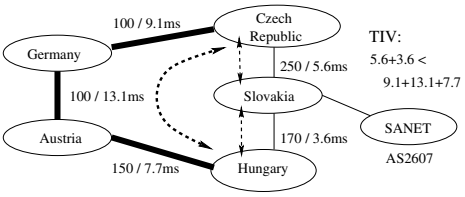

Fig. 7. An example TIV inside the GEANT network

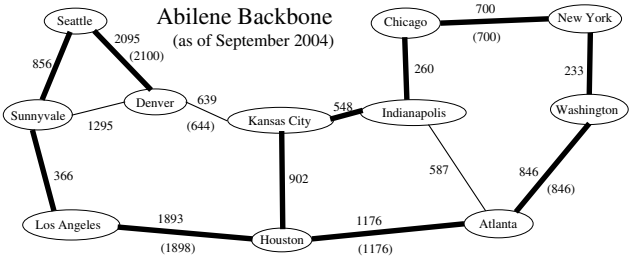

Fig. 8. The topology of Abilene, an Internet2 high-performance backbone network. Thick lines represent OC-192 10Gbps fast backbone links, and thin lines represent slower OC-48 links. The numbers on the links represent the IS-IS weights assigned to the links. Where there is a secondary backup link between two nodes, the IS-IS weight of the backup link is shown in brackets

then run the same measurements to collect the minimum RTT data between all pairs of Abilene backbone routers for the same whole week. Each measurement run takes around 8 hours, so we ran the experiments 3 times per day from 12 August 2004 to 18 August 2004, and obtained 21 matrices. There are only 11 backbone routers in Abilene, so each matrix is 11x11 in size. The minimum RTT between each pair of nodes is then taken to compute TIVs. We observed 5 TIVs out of all the 165 triangles. This represents 3.03% TIV inside the Abilene network.

We learned from the Abilene operators that the link weights are assigned according to geographic distance. As geographic distance is in a Euclidean space, we should expect the triangle inequality to always hold. The reality is, however, that even in an ideally designed network like this, TIV can still occur. On close inspection, we can see that all the TIVs are caused by traffic flowing through independent paths between the triple of nodes (e.g. between Indianapolis, Atlanta and Washington). Although geographically the path is shorter, behavior of the intermediate routers (e.g. load, processing delay, priority of traffic, or queuing delay) can affect the end-to-end RTT measurement. However, compared with GEANT, the violations are much less significant in terms of the r metric (defined in [9] as $r = \frac{a}{b+c} * (1 + (a - (b + c)))$, where a, b and c are the three edges of the triangle and a is the longest edge), as shown in Fig. 9.

4.2 Examples of External TIVs

To illustrate the effect of hot-potato routing on TIV, we picked three PlanetLab nodes from JANET (UK), BELNET (Belgium) and NYSERNet (USA), respectively. We run

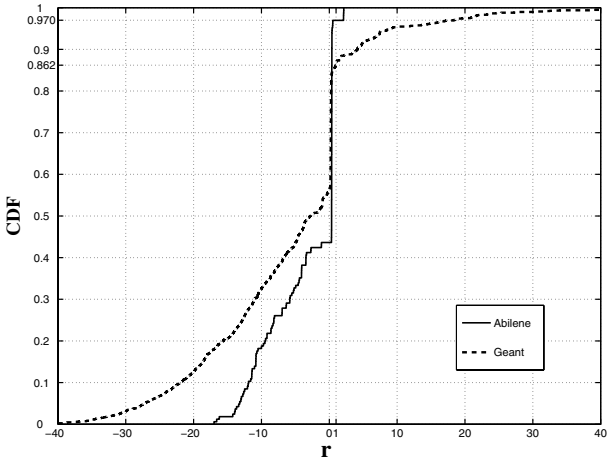

Fig. 9. CDF of the r metric for both Abilene and GEANT. TIVs are signified by $r > 1$, so it can be seen that GEANT exhibits a much higher percentage and magnitude of TIVs

traceroute from each node to the other two nodes to construct the exact path taken by data packets, as shown in Fig. 10. Here we use solid lines to represent a single direct link, and dotted lines to represent a few hops in the middle. We abstract out only the important routers along the paths. We can see that because of hot-potato routing, the traffic from node C to GEANT always goes through the New York router in Abilene. Similarly the GEANT network uses hot-potato routing as well for traffic going to Abilene. The primary link that is causing the problem in this case is the link between NL in GEANT and Chicago in Abilene. This link has a much longer RTT than the NY-to-NY peering link, but is preferred in hot-potato routing to route traffic from NL to Abilene. The end result is that the round-trip path between B and C is asymmetric and much longer than necessary, causing a TIV. We checked that the measurements we obtained with traceroute are within 0.25% error of the minimum RTT value taken during the first week of June 2004 (as mentioned in the dataset in [10]), so we use these RTT measurements in the figure for illustration.

To illustrate the effect of private peering shortcut on TIV, we picked three PlanetLab nodes from JANET (UK), DFN (Germany) and CERNET (China). We run traceroute from each node to the other two to construct the AS-level data path. We then use traceroute to collect RTT data once every 10 minutes between the triple for a 24 hour period on 18 August 2004, and the minimum RTT value is used to demonstrate the TIV. As shown in Fig. 11, the paths between pairs of nodes are symmetric, and there is a private peering shortcut between JANET and CERNET. DFN does not know about this private peering shortcut, so it has to go up the hierarchy tree to communicate with CERNET. This causes a TIV between the three nodes. This AS-level graph corresponds remarkably well with the theoretical analysis shown in Fig. 2.

To illustrate the effect of independent AS paths on TIV, we picked three PlanetLab nodes from Russia, Hong Kong and the UK. The AS-level paths between the three nodes are shown in Fig. 12. Here we ignore the router-level paths within individual ASes, as

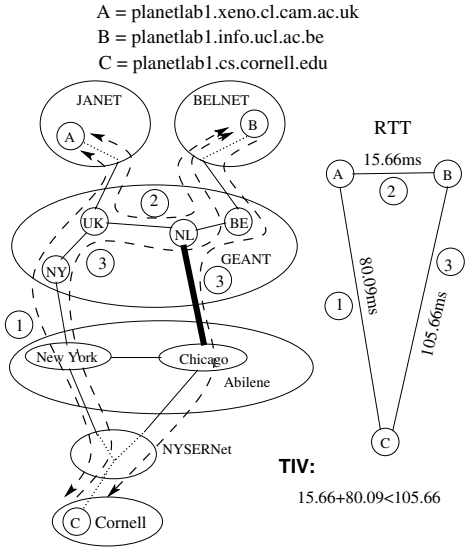

Fig. 10. An example of TIV between PlanetLab sites introduced primarily by hot-potato routing in Abilene, GEANT and NYSERNet

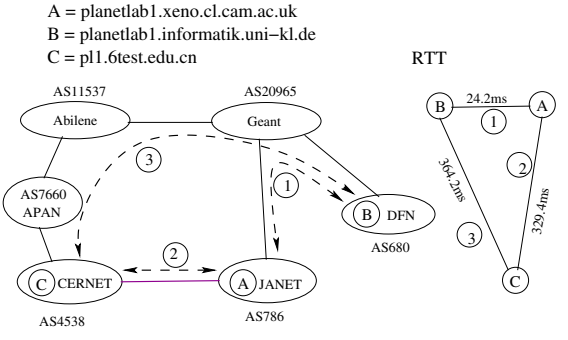

Fig. 11. An example of TIV caused solely by private peering shortcut between ASes, i.e. in this case between JANET and CERNET

these are insignificant. What is of interest in this case is the complicated AS-level paths packets take between these three nodes. By *independent*, we mean that the intermediate ASes are independently engineered and that parts of the AS-paths do not overlap with any other. This is caused by the BGP import and export policies of the ASes involved, and can be explained by the economic incentives of inter-connecting networks [17]. As the minimum RTT measurements between the three nodes vary drastically from week to week in our earlier dataset [10], there are no representative values. However, the common observation is that they all show TIVs between these nodes. So we use the values of a particular measurement in the figure as illustration.

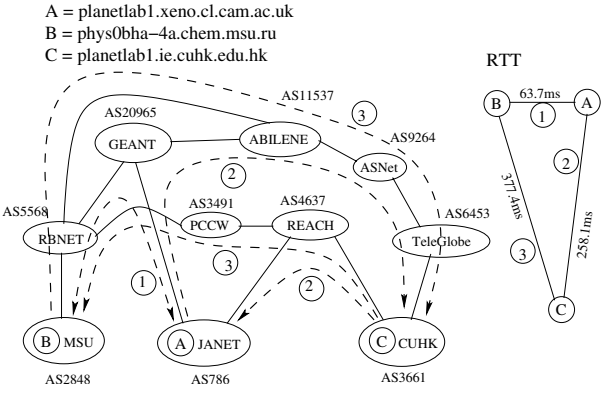

Fig. 12. An example of TIV between PlanetLab sites introduced by independent AS-paths

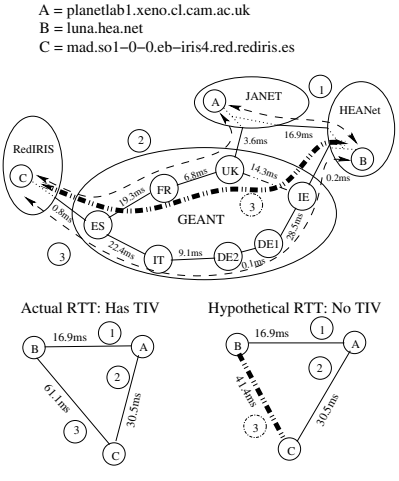

Fig. 13. An example of TIV caused by the interaction between intra-domain and inter-domain routing, or more specifically, between intra-domain TIV and private peering

4.3 Examples of End-to-End TIVs

To show an example of end-to-end TIV caused by the interaction between intra-domain and inter-domain routing, we picked three nodes from JANET (UK), HEANet (Ireland) and RedIRIS (Spain), respectively. As not all the domains have PlanetLab nodes in this case, we use Looking Glass nodes instead whenever necessary. We ran traceroute between the triple, and collected the data once every 10 minutes for a 24 hour period on 18 August 2004. The minimum RTT values between each pair of nodes are then chosen from the dataset. This time, however, we want to see how exactly the path affects TIV, so we use the minimum RTT observation to break the RTT down into segments with one RTT per link (or a set of links if the links are topologically unimportant). The detailed break-down of RTT values is shown in Fig. 13.

Here, we can see that there is actually a link between the UK router in GEANT and the IE router, but its weight is set to be quite high so it is not used in routing. If we used this link on the path from node C to A, then there would not be a TIV between the triple even though there is a private peering link between JANET and HEANet. This illustrates that just looking at the inter-domain structure of the network is not sufficient to determine whether a TIV will occur or not. The behaviors of the intermediate ASes are also important. It is the interaction between intra-domain TIV and private peering shortcut that causes a TIV to occur in this case.

4.4 TIVs in PlanetLab Measurements

To illustrate that TIV is not uncommon in real-world measurements, we took a week's PlanetLab RTT measurement trace from [22] from 13 September 2004 to 19 September 2004. The pair-wise RTT data were collected on consecutive 15-minute periods, and we take the median RTT value from each measurement. Thus for each day in this period there were 96 matrices of RTT measurements, and the size of each matrix is 399×399. Over the week we therefore had 672 such matrices. We then take the minimum RTT value of all the matrices for each pair, and construct a final RTT matrix. Some entries in the final matrix have no values due to unsuccessful measurements, so we denote those by 'NaN'. In calculating triangles, we discard any triangle that has 'NaN' as one of its edges.

We classify all the PlanetLab nodes into research and commercial nodes by looking at whether the IP address of the node matches any prefix in the GREN (i.e. Abilene) BGP table. To be on the safe side, we also manually check the list of nodes that are classified as being commercial nodes. In this way, of all the 399 PlanetLab nodes as of 16 September 2004, we identified 327 nodes as research nodes and 72 nodes as commercial ones. This means that 82% of the hosts are in the GREN, a slight decrease from the 85% we observed in [9].

Notice that the names of the PlanetLab nodes are not accurate indications of whether the node is research or commercial. For example, HP Labs (AS71) has a few PlanetLab nodes under the domain `hpl.hp.com`, but they are reachable from Abilene, and so are in fact research nodes. Conversely, the Computer Science and Artificial Intelligence Lab of MIT has a few PlanetLab nodes under the domain `csail.mit.edu`, but CSAIL (AS40) uses Cogent (AS174) as its upstream provider for connections, and so these nodes are in fact commercial nodes.

We classify all the triangles into 'R.R.R' with all research nodes, 'C.C.C' with all commercial nodes, and 'mixed' with a mixture of nodes. We use the r metric as defined in [9] to illustrate the amount of TIVs. Of all the 2537992 valid triangles formed by the 399 nodes, there were 467328 TIVs, so this represents 18.4% TIV in PlanetLab measurements. Table 1 shows the detailed break-down of TIVs by category for the node-by-node matrix. Fig. 14 shows the CDF distribution of r values for each category, when zoomed in to the area around $r = 1$. We can see that 'R.R.R' triangles behave the best. There are the fewest number of TIVs, and the TIVs are all quite small in magnitudes. 'C.C.C' and 'mixed' triangles behave very similarly, and both have a higher percentage of TIVs and much bigger r values.

Table 1. A detailed break-down of TIVs by category for the node-by-node matrix

Category	Total	TIVs	Percentage
$R.R.R$	1704809	282164	16.6%
$C.C.C$	9678	2306	23.8%
$Mixed$	823505	182858	22.2%
All	2537992	467328	18.4%

Table 2. A detailed break-down of TIVs by category for the site-by-site matrix

Category	Total	TIVs	Percentage
$R.R.R$	263062	52170	19.8%
$C.C.C$	966	227	23.5%
$Mixed$	119945	26084	21.8%
All	383973	78481	20.4%

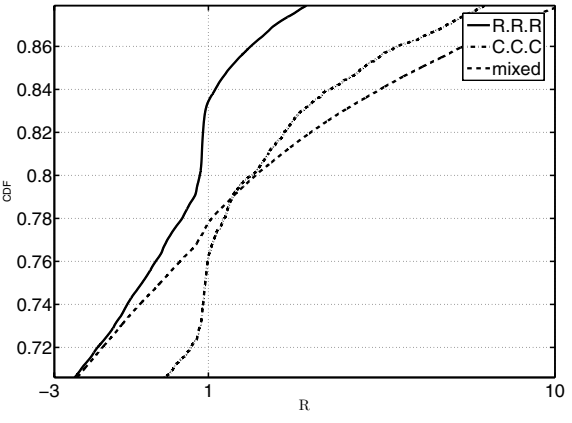

Fig. 14. The CDF distributions of r values for the three categories of triangles formed by PlanetLab *nodes*, when zoomed in to areas around $r = 1$

However, there are a few problems with using the node-by-node matrix. When we look at the r values close to 1, we can see that most of those corresponding triangles have two nodes physically located in the same site with a very small RTT value between them (typically $<$1ms). This makes the r value very vulnerable to measurement artifacts, as measurement error can often be much bigger than 1ms. So if we do not filter out those triangles, then our measured TIV percentage is not accurate. We have also seen some absurd RTT values (e.g. $>$3000ms), and triangles with one edge having those RTT values usually have huge r values. We suspect the high RTT values are due to a few nodes being overloaded, so we want to filter out those as well to get an accurate measure of TIV magnitude. For these reasons, we want to get a site-by-site matrix, where we pick a representative node for each site.

In our initial approach, we noted that all the nodes that belong to a physical site are located within the same '/24' prefix, so we picked one node from each '/24' prefix which has the most number of non-NaN measurements that are less than 1500ms. However, we also noticed that sometimes not all the nodes in the same '/24' prefix are physically located together. For example, the Internet2 PlanetLab nodes are physically scattered across the Abilene backbone and co-located with Abilene backbone routers, and so the RTTs between them are typically more than 10ms. Therefore, we decided to use physical closeness and the *behavior* of the nodes as the main criteria for grouping nodes into sites.

For each node (i.e. the *pivot* node), we first pick out all the nodes that have sub-millisecond RTT to that node. Within the group, we then check each pair-wise RTT to make sure it is either sub-millisecond or NaN. Then, we calculate the Correlation Coefficient (CC) between the *pivot* node and every other node, and throw away nodes that have $CC < 0.99$. In calculating CC, we use the standard definition on the two row vectors of the RTT matrix, but we modify it slightly so we treat NaN entries as perfect matches with the corresponding entries. Finally, for each remaining node in the group, we also search through the node list and add into the group any node that has the same '/24' prefix as this node but NaN in the RTT matrix. This accounts for nodes in the same physical location but was down during the measurement period. By using the above procedure, we were able to reduce the 399 nodes into 168 'sites', with 140 research 'sites' and 28 commercial 'sites'. Again we picked out the 'best' node from each site, and calculate TIVs on the reduced 168×168 matrix. This time, we observed 78481 TIVs out of all 383973 valid triangles. Table 2 shows the detailed break-down of TIVs by category for the site-by-site matrix.

We expected the 'R.R.R' triangles to behave the best, as it is very likely that traffic between them is only transited through GREN. In GREN, the networks are not driven purely by commercial relationships and are very cooperative in general, and the GREN inter-domain routing policies are often configured to use shortest path even if it violates economic provider-to-customer relationship. For example, RENATER2 (AS2200), the French research network, has a direct peering connection with APAN-Korea (AS9270), constructed under the TEIN2 project. Although RENATER2 is a customer of GEANT (AS20965), it still exports this private peering link to GEANT so the shortest path can be used. (Actually the European Commission is funding part of the TEIN2 project, so this private peering is logically peering with both RENATER2 and GEANT. GEANT can also reach APAN-Korea through Abilene and APAN-Japan, but the path is much longer.) In the commercial world, however, paths are often inflated due to economic reasons [23,24]. Thus, we would expect a higher percentage of TIVs between commercial PlanetLab nodes than between research ones. When there is a mixture of nodes, traffic between a research and a commercial node tends to go through the commercial Internet, so in a mixed triangle there are at least two paths between nodes through the commercial Internet. This makes the mixed case a lot like the 'C.C.C' case.

5 Conclusion

Although Internet routing policies play an important role in the global observed round-trip-times, we also want to emphasize that routing policy is not the only thing that contributes to TIVs. We have already seen that the private peering connection between JANET and CERNET is not giving too much savings on the RTT measurements. This relates to the Layer 2 technology used on this peering link. There is also the fact that the earth is a sphere, not a plane, so triangles on the surface of earth can go around the earth and do not have to satisfy the triangle inequality (although currently there are only very few fast links through continental Europe to Asia). Even as we are finishing this paper, we have heard that the South American research network is being re-structured and connected to GEANT directly, and the AMPATH network, which used to connect Brazil

to Abilene, is being decommissioned. This would mean that temporarily all research traffic from Brazil to Abilene will need to go through GEANT, essentially traversing through the Atlantic twice. In particular, we have verified that traffic from Brazil to Mexico goes along this very much stretched path through GEANT and Abilene. This illustrates that the structural planning of the network can bring about unexpected TIVs as well.

In conclusion, TIVs are not just data collection artifacts, they can be structurally persistent as a result of routing policies (as well as many other factors). Both intra-domain routing policies and inter-domain routing policies, and the interactions between them, can cause structural TIVs. TIVs present an opportunity for overlay routing, but they make Internet Coordinate embeddings less accurate.

Acknowledgements

The bulk of the work done for this paper was carried out while the authors were at Intel Research, Cambridge. The authors would like to thank Derek McAuley of Intel Research for his support. In addition, Jon Crowcroft of the University of Cambridge provided many useful comments and suggestions.

References

1. Andersen, D.G., Balakrishnan, H., Kaashoek, M.F., Morris, R.: Resilient overlay networks. In: Proc. 18th ACM SOSP, Banff, Canada. (2001)
2. Ng, T.E., Zhang, H.: Predicting Internet Network Distance with Coordinates-Based Approaches. In: IEEE INFOCOM' 02, New York, USA (2002)
3. Pias, M., Crowcroft, J., Wilbur, S., Harris, T., Bhatti, S.: Lighthouses for Scalable Distributed Location. In: 2nd International Workshop on Peer-to-Peer Systems. (2003)
4. Tang, L., Crovella, M.: Virtual Landmarks for the Internet. In: ACM SIGCOMM Internet Measurement Conference (IMC'03), Miami (FL), USA (2003)
5. Lim, H., Hou, J., Choi, C.: Constructing Internet Coordinate System Based on Delay Measurement. In: ACM SIGCOMM Internet Measurement Conference (IMC'03), USA (2003)
6. Costa, M., Castro, M., Rowstron, A., Key, P.: PIC: Practical Internet Coordinates for Distance Estimation. In: 24th IEEE International Conference on Distributed Computing Systems (ICDCS' 04), Tokyo, Japan (2004)
7. Dabek, F., Cox, R., Kaashoek, F., Morris, R.: Vivaldi: A Decentralized Network Coordinate System. In: Proceedings of the ACM SIGCOMM 2004, Portland, Oregon (2004)
8. Shavitt, Y., Tankel, T.: Big-bang simulation for embedding network distances in Euclidean space. In: Proceedings of the IEEE INFOCOM 2003, San Francisco, California (2003)
9. Banerjee, S., Griffin, T.G., Pias, M.: The Interdomain Connectivity of PlanetLab Nodes. In: 5th International Workshop on Passive and Active Measurement, France (2004)
10. Pias, M., Lua, E.K., Zheng, H., Griffin, T.G.: On the Accuracy of Embeddings for Internet Coordinate Systems. In: Under submission. (2005)
11. Shavitt, Y., Tankel, T.: On the Curvature of the Internet and its usage for Overlay Construction and Distance Estimation. In: Proc. ACM INFOCOM 2004, Hong Kong (2004)
12. Perlman, R.: Interconnections: Bridges, Routers, Switches, and Internetworking Protocols. 2nd edn. Addison-Wesley (1999)
13. Pepelnjak, I.: EIGRP Network Design Solutions: The Definitive Resource for EIGRP Design, Deployment, and Operation. Cisco Press (2000)

14. Greenberg, A.G., Hajek, B.: Deflection routing in hypercube networks. IEEE Trans. Commun. **40** (1992) 1070–1081
15. Norton, W.B.: Internet Service Providers and Peering. In: Proceedings of NANOG 19, Albuquerque, New Mexico (2000)
16. Bu, T., Gao, L., Towsley, D.: On routing table growth. In: Proc. of Global Internet Symposium 2002. (2002)
17. Houston, G.I.: Interconnection, Peering and Settlements. Internet Protocol Journal (1999)
18. Abilene: Abilene Network Operations Center: Conditions of Use (COU) for the Abilene Network. http://abilene.internet2.edu/policies/cou.html (2004)
19. DANTE: GEANT Topology. http://www.dante.net/server/show/nav.007009007 (2004)
20. DANTE: GEANT Backbone Looking Glass. http://stats.geant.net/lg/lgform.cgi (2004)
21. Abilene: Abilene Network Operations Center: Abilene Backbone Network Router Configurations. http://loadrunner.uits.iu.edu/~gcbrowni/Abilene/vn/configs/configs.html (2004)
22. Jeremy Stribling: Pair-wise RTT data between all PlanetLab nodes by Jeremy Stribling (MIT). http://www.pdos.lcs.mit.edu/~strib/pl_app/ (2004)
23. Spring, N., Mahajan, R., Anderson, T.: Quantifying the Causes of Path Inflation. In: Proceedings of ACM SIGCOMM 2003. (2003)
24. Gao, L., Wang, F.: The Extent of AS Path Inflation by Routing Policies. In: Proceedings of Global Internet 2002. (2002)