Computer Communications 32 (2009) 913–926

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

An investigation of the Internet's IP-layer connectivity

Li Tang [a,b,d,*], Jun Li [b,c], Yanda Li [a,c], Scott Shenker [d]

[a] Department of Automation, Tsinghua University, Beijing,China
[b] Research Institute of Information Technology, Tsinghua University, Beijing, China
[c] Tsinghua National Lab for Information Science and Technology, Beijing, China
[d] ICSI, 1947 Center Street, Suite 600, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 11 March 2008
Received in revised form 17 December 2008
Accepted 21 December 2008
Available online 31 December 2008

Keywords:
Internet
Connectivity
Reachability
Reliability
Routing dynamics

ABSTRACT

The Internet's tremendous value is undoubtedly dependent on its universal connectivity among a great number of heterogeneous networks that are distributed over the world. In recent years, while the Internet's scale has expanded exponentially, the current status of its connectivity is still in the lack of comprehensive and formal study. In this paper, we contribute to the understanding of Internet's IP-layer connectivity by quantitatively measuring the reachability from 124 PlanetLab nodes towards 197869 diversely distributed destination IP addresses. We first demonstrate our methodology to meet the challenges in experiment design, and then statistically analyze the Internet's IP-layer connectivity in various aspects, including the directly reachable proportion, packet loss, delay variation, and the effect of domain and geographic distance. Finally, we investigate main causes of IP-layer unreachability by revealing some intentional insulation policies based on empirical study on a few special cases and analyzing its correlation to typical routing issues.

1. Introduction

As the Internet has been increasingly used as a universal information and communication platform in recent years, its reliability is accordingly becoming more and more important. Unfortunately, however, due to its scale expansion and complexity increase, the IP routing infrastructure, which is the substructure of all Internet applications and services, tends to suffer even increased instability and unavailability. It is well-known that a wide variety of problems can cause IP-layer end-to-end (E2E) path failures, including router mis-configurations, network maintenance, physical device breakdown, and routing misbehaviors. Therefore, as the Internet is evolving to support critical applications, such as online banking and medical treatment, it urges the demand of investigating the current Internet's IP-layer reliability.

Previous research has used availability as a metric to study the Internet's reliability [1–3]. Availability is usually defined as the fraction of time that a service is reachable and correctly functioning relative to the time it is in operation. As the measurement of availability requires sending probing messages rather frequently in order to monitor continuous reachability, this approach has to limit the experiment in a relatively small set of targets so as not to trigger too much measurement traffic.

In this paper, we contribute to the understanding of the current Internet's IP-layer reliability from a novel perspective. Specifically, given a set of vantage points, we aim to quantitatively measure their reachability to the rest part of the Internet. We refer to this new metric as *connectivity* to distinguish it from the previous *availability* metric. While *availability* mainly reveals the probability of a service correctly functioning when it is accessed at different time, *connectivity* reflects the likelihood of its correctly functioning when being accessed from different locations. The distinction between *availability* and *connectivity* is important in the sense that while *availability* has been widely studied in literature [1–4], *connectivity* has not. Unlike *availability* merely subjected to unexpected failures, *connectivity* can also be limited by intentional insulation for business and security purposes. In addition, the significance of a comprehensive study on the Internet's IP-layer connectivity has also been emphasized by the fact that thanks to the prevalence of wireless networks, more and more people start accessing to the Internet and online services from many different locations, including not only private places such as home and offices but also some public locales such as hotels and libraries. Accordingly, a critical application on the Internet should provide reliable service not only at most time, but also accessible from a vast diversity of marginal access networks that are geographically located in different places.

In the rest of this paper, we first present the methodology used in our experiment design to efficiently measure the Internet's IP-layer connectivity (Section 2), and then we show to what extent our measurement dataset is representative of the common case in the whole Internet (Section 3). With all the measurements

* Corresponding author. Address: Room 3-421, FIT Building, Tsinghua University, Beijing 100084, China. Tel.: + 86 13810394042.
E-mail address: tangli03@mails.tsinghua.edu.cn (L. Tang).

0140-3664/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2008.12.015

collected from 124 vantage points, we begin with an overview of the current situation of the Internet's IP-layer connectivity (Section 4). Afterwards, we statistically analyze the measurements from 118 typical vantage points in a variety of aspects, including the directly reachable proportion, packet loss, delay variation, and the effect of domain and geographic distance (Section 5). Next, we empirically study the strikingly high IP-layer unreachability from the other 6 special vantage points, revealing several unusual insulation policies set intentionally by networks operators (Section 6). Finally, we analyze how IP-layer unreachability is related to the Internet's typical routing issues (Section 7) and summarize our main observations (Section 8).

2. Experiment methodology

Our study framework is to measure a typical sample of IP-layer E2E reachability between a set of topologically diverse hosts on the Internet. We expect the measurements to be plausibly representative enough for us to gain quantitative insight into a considerably rich cross-section of the current Internet's IP-layer connectivity. While this idea is conceptually simple at the first glance, it actually requires to solve the following difficult challenges in practice.

2.1. Selection of vantage points

First, it requires a group of suitable vantage points to measure and collect data that can hopefully characterize the Internet's heterogeneous feature and represent its topological diversity. For this purpose, we selected 278 PlanetLab nodes located in different sites to be the vantage points, each of which was suggested by CoMon project [5] to be the best node out of its site.

2.2. Selection of targets

Second, it also requires an appropriate set of targets that can be persistently on-line and topologically representative. Although Web servers are mostly persistently on-line, they are not good candidates for our study, because resolving the Web server's names involves a series of complex DNS recursive operations. In addiction, in order to improve the scalability and availability, large Web content providers have widely utilized mirroring, content distribution, and ISP multi-homing techniques, which enable them to adaptively return different IP addresses to the same DNS name [6]. The possible failure and uncertainty of DNS operations will inevitably interfere with the evaluation of IP-layer reachability.

On the other hand, given that around 2985 million IPv4 addresses have been allocated so far [7], a complete measurement of reachability to every address is impractical either. For one thing, such an experiment would cause huge measurement traffic and take a long time to finish. For another, most targets would probably turn out to be unreachable just because they are absence/offline rather than interrupted by failures of the Internet's IP-layer routing service. Thus, an efficient yet reasonable clustering method is demanded for selecting typical online targets.

To solve this problem, we leveraged the measurement results shared by iPlane project [8]. iPlane performs `traceroute` towards a carefully chosen target list of IP addresses, each of which is located in a specific BGP atom [9]. A BGP atom is a set of globally routable IP prefixes, each of which has the same AS path to it from any given vantage point according to the observed BGP routing table snapshots. iPlane's target list achieves both measurement efficiency and wide topology coverage. To further reduce the effect of absent/offline targets, we picked out only the reachable IP addresses out of iPlane's `traceroute` results archived in several continuous days as the final targets in our experiment. Specifically, we found that the time period of three days is a good choice to balance the information's completeness and freshness. It is long enough for iPlane to finish a round of complete probe towards all its targets, and yet most online targets found would hopefully continue to be online in the near future. Section 3.1 shows the completeness and distribution feature of the targets in more details.

2.3. Design of probe method

At last, it requires an effective reachability probe method. We used the most popular IP-layer troubleshooting utility `ping` to test E2E reachability from each vantage point to every target. On probing a target, a vantage point sent five ICMP ECHO_REQUEST packets with an interval of one second to the target. The probe ended up either after the vantage point had obtained five ICMP ECHO_REPLY packets or until it timed out by ten seconds. In the former case, the vantage point continued to probe the next target. In the latter case, more measures would be carried out to infer the reasons of unreachability. Specifically, the vantage point would delegate the probe task towards the same target to another randomly selected vantage point by means of secure shell (SSH) tunnels. The delegation attempt would be repeated at most ten times before the target had been successfully reached or until five other vantage points all failed to reach the target. If the target turned out to be reachable by some other vantage point, we attributed the direct unreachability between the origin vantage point and the target to backbone failures; otherwise, to edge network failures of either the original vantage point's or the target's access networks. Finally, after the delegation process, if any, the original vantage point would directly probe the target again, which we refer to as confirmation process.

2.4. Debate of shortcomings

A legitimate shortcoming of the methodology is the lack of ability to identify the specific causes leading to IP-layer direct unreachability, which is actually inherent in all E2E measurement techniques across the Internet. While E2E measuring can efficiently reflect a quantity of direct interest to network end users, it has difficulty in localizing a problem out of a compound of effects at different hops in the networks.

One way to obtain more detailed information is to replace ping with `traceroute` as the probe method. However, because `traceroute` sends much more probe packets and takes longer to finish a measurement, it is not as scalable as `ping`. As our experiment needs to probe a vast number of targets, it cannot afford frequently `tracerouting` every target multiple times. With just one `traceroute` result per target, it is difficult to quantitatively estimate and diminish the effect of routing changes that possibly occur during the `traceroute` measurement. Moreover, as shown in Section 6.2, some security configuration against port scanning attacks may prevent `traceroute` from reaching the target even when the target can be successfully reached by `ping`.

Another shortcoming is that our probe method using `ping` was based on ICMP packets which, compared to TCP and UDP packets, were possibly treated in different ways due to security and routing policies. While almost all Internet applications use either TCP or UDP as their transport-layer protocols, ICMP protocol is originally designed to maintain, control and diagnose IP network with out-of-band messages. Although ICMP is a required protocol tightly integrated with IP and all routers are expected to be compatible with it [10], ISP operators still have incentive to restrict certain types of ICMP packets by setting up relevant policies due to security concerns or concealment of inside topologies. In Section 6.2, we indeed reveal some detailed policies that particularly forbid certain types of ICMP packets based on empirical study. On the other hand, however, we argue that such special polices do not

distort our study results from the real situations. If the IP-layer unreachability between a pair of vantage point and a target observed in our experiment was caused by some special policies forbidding the ICMP packet utilized by `ping`, then either the vantage point could not reach an extremely high percentage of targets (if the policies were set by the vantage point's domain), or every vantage point all could not reach the target (if the policies were set by the target's domain). In the former case, the vantage point is taken as a special case in this paper and studied in particular in Section 6. In the latter case, the effect of this target can be eliminated by the adjustment methods proposed in Section 4.2. Therefore, using `ping` and ICMP in our experiment actually has negligible influence to our statistical analysis of the typical cases.

3. Dataset representativeness

Our experiment finally collected about two hundred million measurements of Internet's IP-layer reachability, from 124 vantage points against 197869 targets. While we do not claim these data give a complete view of the Internet's IP-layer connectivity, we do argue that they can reveal a considerably representative cross-section of its current situation. In the rest of this section, we explicate such representativeness in details.

3.1. Target coverage

We totally picked out 197869 reachable targets out of the archives collected by iPlane project on the end of 2007. It is acknowledged that there is no coordinate mechanism embedded in the current Internet's architecture, and thus it is impossible to obtain a complete map of the Internet's topology. Given this, to examine the representativeness of these targets related to the whole Internet, we investigated their topological coverage on three aspects, autonomous system (AS) level, point of presence (PoP) level, and geographic location level. In the rest of this subsection, before presenting the final results, we first explain the meaning of each aspect, the methods used for inferring and mapping the target's IP addresses to their corresponding AS numbers (ASN), PoP numbers (PoPN) and city names, and the analysis on the limitation of these methods.

The concept of AS is introduced due to the increasing requirements of scalability, commercialization, and privatization to the Internet. An AS is an independently organized and operated network or collection of networks, which interconnected with each other compose the current Internet's routing substrate. Examples of AS operators include universities, large businesses, Internet service providers (ISP), and telephone companies. Currently, there are mainly two ways to map an IP address to its corresponding AS number. One way is based on WHOIS databases that are manually maintained by Regional Internet Registries [11]. While the databases contain a wide range of information of the allocated IP addresses, there is little requirement for updating the registered information in a timely fashion. Considering that ISPs are constantly changing their topologies and traffic policies, we chose the other way that uses more seasonable information collected from operant routers running the de facto inter-domain routing protocol BGP (currently in its fourth version). A BGP routing table contains all the globally routable IP prefixes, each of which is followed by a series of ASes indicating the AS path to that prefix. The iPlane project has extracted BGP routing information into two dictionaries. In the first dictionary, iPlane merges BGP routing tables from several sources such as RouteViews [12] and RIPE [13] to assemble a large set of AS paths, and deems the origin AS of an IP prefix corresponding to each AS path to be the last AS on that path. In case that multiple ASes correspond to the same IP prefix, we remained only one of the ASes for simplicity. While the origin-AS-mapping dictionary is indexed by IP prefixes, the other one, referred to as the exact-IP-to-AS-mapping dictionary, is indexed by specific IP addresses. The second dictionary is obtained by further revising a router's every interface to be the AS that accounts for a majority among the origin ASes of the interface's aliases. To obtain the ASN for a target, we first look up the target's IP address through the exact-IP-to-AS-mapping dictionary by precise matching, and then through the origin-AS-mapping dictionary by longest prefix matching only if there is no valid ASN returned by the first lookup.

An AS's inner topology consists of its backbone networks and PoPs. Each PoP is a physical location (usually in the grain of a city) where the AS houses a collection of routers. High-speed optic fibers are used to connect different PoPs composing the backbone networks. The iPlane project provides an IP-to-PoP-mapping dictionary, in which all IP addresses that are in the same AS and located in the same location are clustered together and mapped to an identical PoPN. Unfortunately, however, with this dictionary, we could only obtain the PoP-level information of partial targets. Only 110248 out of all 197869 targets had their PoPNs, and among them only 32572 targets had valid latitude and longitude values of their PoPs. To solve this problem, we attributed each IP address's PoP-level information in the IP-to-PoP-mapping dictionary to the IP address's longest matching IP prefix in the origin-AS-mapping dictionary. In this way, a target failing to get its exact matching PoP-level information from the IP-to-PoP-mapping dictionary could further attempt to get its longest prefix matching result. Thanks to this mechanism, we managed to increase the number of targets that have PoP-level information by about 40% to 154521, among which 45933 have valid latitude and longitude values.

We attributed a target's geographic location to the city where its corresponding PoP located. We used a DNS-based-location dictionary provided by iPlane project to infer the location of each PoP. To generate this dictionary, iPlane first obtains the DNS names of the routers possessing corresponding IP addresses to a PoP, and then interprets the DNS names with rules from the undns [14] and sarangworld [15] projects. As there were some PoPs in the dictionary having only latitude and longitude values but no corresponding city name, or the other way around, we referred to GeoWorldMap's city data [16] to complement the lacked counterparts.

Combining all efforts above, we were able to construct a comprehensive dictionary that maps a target's IP address to its ASN, PoPN, and geographic location that includes the latitude and longitude values as well as the city and country names. Table 1 summarizes the coverage of our target list used in this paper relative to iPlane's dictionaries and the whole Internet, respectively, in terms of the numbers of IP prefixes, ASes, PoPs, and cities/countries. The data about the whole Internet were obtained from the timely BGP report on the active BGP entries and advertised AS numbers [17], and only serve as a rough reference.

3.2. Vantage point distribution

We eventually managed to collect probe measurements from 124 vantage points that had successfully finished probing all the

Table 1
The coverage of chosen targets relative to those observed by iPlane and in the whole Internet.

Scope	IP prefix	AS	PoP	City/Country
In target list	20379	19932	44118	911/69
Observed by iPlane	247608	25610	140598	1702/74
Total in the Internet	249365	26917	Unknown	Unknown

Table 2
The topological coverage of vantage points.

IP prefix	AS	PoP	PlanetLab site
116	95	121	124

chosen targets. The distribution information of these vantage points came from two sources. The dominant source was the vantage point's official registration information on PlanetLab, including its name, IP address, site's name, and the latitude and longitude values of its site. Besides this source, we also inferred each vantage point's ASN, PoPN, and geographic location according to its IP address, in the same way as we processed the targets in the last subsection. The results indicated that the inferred geographic locations were mostly consistent with the official registrations. For a few other vantage points, neither of the above two sources provided their locations and we manually found and added their geographic information according to their DNS names. We summarize the topological coverage of the vantage points in Table 2, and illustrate their detailed geographic distribution in Fig. 1.

4. Overview of Internet connectivity

In this section, we present an overview of all the measurements in our data collection, and then we introduce the framework used in our further analysis.

4.1. Original measurement results

To give an overview, we classify the measurements towards all the targets into five categories. According to the probe method described in Section 2.3, a given vantage point attempted at most three processes to reach each target, namely the first round of direct probe that consists of five ICMP ECHO_REQUEST packets, the delegation process that makes use of one of at most five other vantage points, and the confirmation process by performing another round of direct probe. We term category '*C*1' for the case that the target was reachable by the vantage point immediately in the first round of direct probe; '*C*2' for the case that the target was unreachable in the first round of direct probe, successfully reached by using another vantage point during the delegation process, and still unreachable in the confirmation process; '*C*3' for the case that except the first round of direct probe, the target was reachable in both the delegation and confirmation processes; '*C*4' for the case that the target was unreachable in all three processes; and '*C*5' for the case that the target was reachable only in the confirmation process, but not in the first round of direct probe nor delegation process. For ease of expression, we will also refer to the case in category *C*4 with the term'*complete unreachability*' to indicate the meaning that in all three processes every relevant vantage point all failed to reach the given target.

Generally speaking, the percentage of category *C*1 manifests the likelihood that a vantage point was reliably reachable by means of the Internet's IP-layer routing service from a variety of different access networks and locations, and the percentage of category *C*4 manifests the lower bound of the likelihood that the vantage point was unreachable due to critical IP-layer problems, which most probably happened in the vantage point's or target's access networks. In contrast, the interpretation of category *C*2, *C*3 and *C*5 is relatively empirical and complicated. In category *C*2, as the IP-layer connectivity between the original vantage point and the target was interrupted in both the first direct probe and confirmation processes and the interrupt could be detoured by using another vantage point in the delegation process, it is likely that the IP-layer unreachability was caused by some long-term failures in backbone networks. As regard to category *C*3 and *C*5, the IP-layer connectivity being interrupted only in the first direct probe process but recovered in the confirmation process implies that the IP-layer unreachability between the original vantage point and the target was caused by either a short-term congestion or a long-term failure just recovered during the measurement. Moreover, while the congestion or failure in category *C*3 was likely to happen in backbone networks, that in *C*5 is more likely in edge networks, because the interruption in category *C*3 could be detoured in the delegation process, but that in *C*5 could not.

Fig. 2 illustrates the cumulative distribution function (CDF) plots of each category's percentages from every vantage point. As can be seen, except a few particular cases, most vantage points have very similar percentages of the same category. There are totally 6 vantage points whose category *C*1's percentages are smaller than 0.9, the knee of *C*1's CDF plot in Fig. 2. We take these 6 vantage points as special cases, and postpone related discussion about them to Section 6. In the next section, we base our statistical analysis on the measurements from the other typical vantage points. The average percentage of each category among the remained typical vantage points is given in the second row of Table 3.

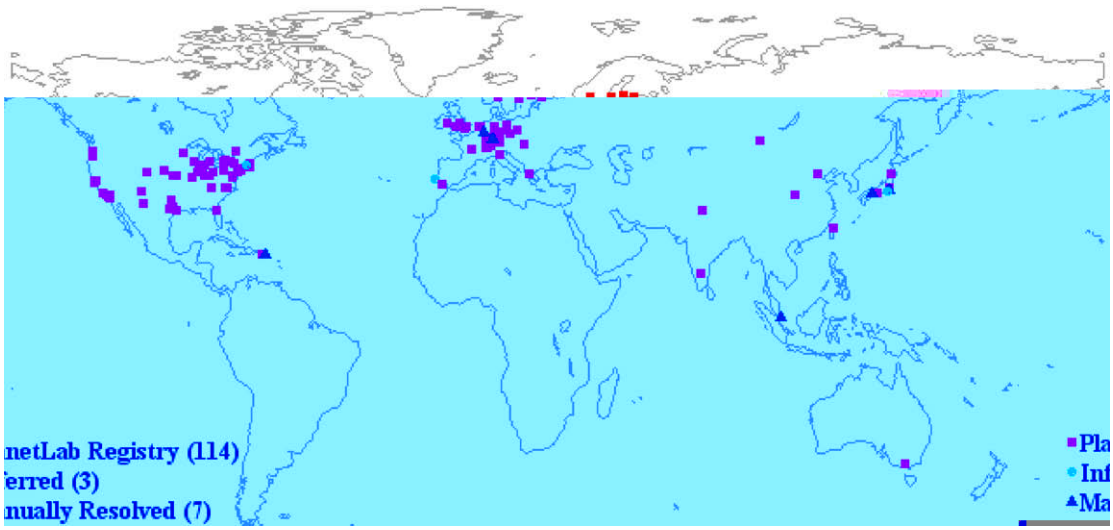

Fig. 1. The geographic distribution of the vantage points.

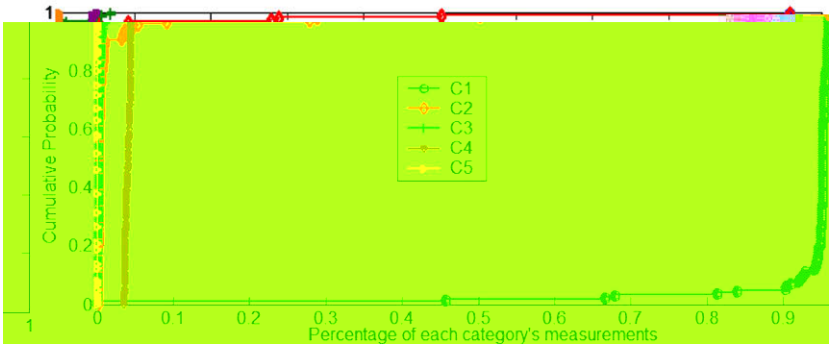

Fig. 2. The CDF plots of each category's percentage from every vantage point.

Table 3
The topological coverage of vantage points.

Category	$C1$ (%)	$C2$ (%)	$C3$ (%)	$C4$ (%)	$C5$ (%)
Raw measurements	95.01	0.59	0.33	4.05	0.02
Conservative adjustment	95.63	0.60	0.33	3.42	0.02
Majority adjustment($k = 60$)	98.08	0.61	0.34	0.95	0.02
NVP-proportion adjustment	97.87	0.60	0.33	1.26	0.02

4.2. Adjustment methods

Although we had tried our best in the experiment design to select targets that were likely to be online, the final target list might still contain some targets whose corresponding end hosts were absent/offline during the period of our measurement. In this sense, the category of $C4$ actually consists of two parts. We use $P1$ to denote the first part in which complete unreachability was caused by the outage of the Internet's IP-layer routing service and $P2$ to denote the second part in which complete unreachability was cause by the absence/offline of the corresponding target. While $P1$ is what we really expect to quantify in our study, $P2$ is actually interference that can lead to an overestimation of the percentage of category $C4$ and thus to an underestimation of the current Internet's IP-layer connectivity. Consequently, it necessitates some adjustment to alleviate this effect.

Intuitively, if a target was completely unreachable by many vantage points at different times, the cause of its unreachability was then much more probably due to the absence/offline of the target, rather than IP-layer outage. To gain deeper insight on this issue, for each one of the 24506 targets contained in $C4$, we count the number of vantage points that are completely unreachable to the target, and illustrated the CDF plot (NVP) in Fig. 3. As can be seen, there are indeed 1295 targets completely unreachable by all the 118 typical vantage points. On the other hand, however, these targets are only negligible 0.65% of all the 197869 selected targets, indicating pretty good effectiveness of our target selection method. Fig. 3 also shows that the number of vantage points completely unreachable to the same target, which we refer to as the target's NVP for brief, can vary in a wide range. As a result, it is obscure to define how large a target's NVP should be considered as the sufficient condition to judge that the target was surely absence/offline. By judging a target to be offline if it has a larger NVP than a given threshold k, we show in Fig. 3 (PUT) how the proportion of $P2$ relative to the whole category $C4$ varies as k is assigned different values. Given this, we explore two adjustment strategies: one is the conservative adjustment which assigns k to be the largest possible value 118, and the other is to believe in the majority's opinion and let k be 60.

Unfortunately, besides the difficulty in determining a suitable value of k, another shortcoming of the above adjustment methods by setting an NVP threshold is that the measurements from all vantage points towards the same target are either absolutely classified into $P1$ or absolutely into $P2$. In fact, however, because each vantage point measured its reachability to the target asynchronously, it is probable that some vantage point's complete unreachability to a target was caused by IP-layer outage and their measurements should be classified into $P1$, while other vantage point's complete unreachability to the same target was caused by the target's absence/offline and their measurements should be separately classified into $P2$. Taking this issue into consideration, we propose another adjustment method that classifies each of category $C4$'s measurements into $P2$ with a probability that is in proportion to the measurement's target's NVP. In other words, if a vantage point was completely unable to reach a target that has its NVP of n, we attribute the reason of this complete unreachability to the target's absence/offline with a probability of $\frac{n}{max(NVP)}$. With this adjustment

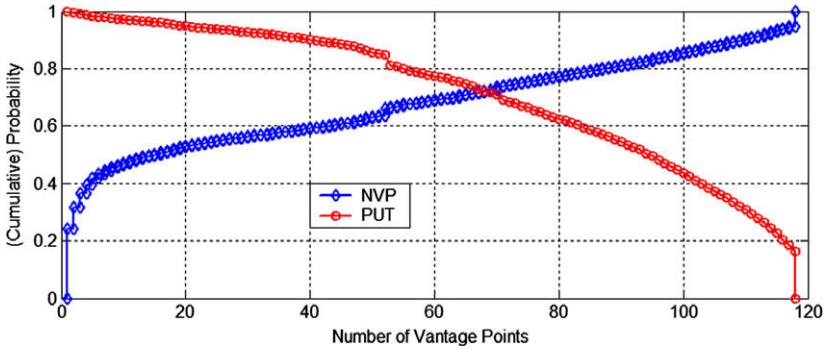

Fig. 3. The CDF plots of the number of vantage points unable to reach each target in category $C4$.

method, we find that 72.02% of category $C4$'s measurements are classified into $P2$, which is similar to the results obtained by the NVP-threshold-based adjustment method with $k = 68$. Table 3 presents each category's average percentage after using different adjustment methods.

4.3. Preliminary analysis

Unsurprisingly, the results show that the connectivity of the IP-layer routing service over a wide coverage of the Internet is a little smaller than the availability of critical Internet's infrastructures, such as backbone routers and Web servers. Labovitz et al.'s work showed that the overall uptime of Internet's backbone routers averaged above 99.0% [18], and Dahlin et al found that the availability of Web servers were roughly ranged between 98.1% and 99.3% [19]. In contrast, the IP-layer connectivity turns out to be 93.2% on average without any adjustment, and even if adjusting category $C4$ with the most radical strategy, the IP-layer connectivity is still no more than 98.08%. In addition to the different significance of the systems, another probable reason is the exponential scale expansion of the Internet in recent years [20]. It implies that despite the improvement of routing protocols and hardware devices, it is still a stringent challenge to fit the fast growth of the Internet without degrading its reliability.

Comparing category $C4$'s percentage to the sum of category $C2$'s and $C3$'s percentages in Table 3, we observe that majority Internet's IP-layer unreachability was caused by the failure of edge networks or end hosts, rather than that of the Internet's backbone's. As a result, only around 25–50% of the unreachability could be recovered by using a proper relay node to detour the IP-layer routing outage, and only around 15–30% could be recovered by retrying the IP-layer routing service again after a short period of time (from dozens of seconds to no more than several minutes). We have noted the possibility that our experiment implementation can limit the effectiveness of these detouring and retrying schemes for recovering Internet's IP-layer unreachability. For example, had a richer and larger set of vantage points been attempted in the delegation process, maybe the detouring scheme would be able to recover more IP-layer unreachability. Although this implementation limitation is impossible to prevent completely, we find that its effect is actually negligible to the results of our study. We will revisit this issue in details in the following section.

5. Statistical analysis of typical cases

In this section, we analyze the measurements collected from 118 typical vantage points in many different aspects, and aim to reveal a statistical perspective of the current Internet's IP-layer connectivity.

5.1. Packet loss

5.1.1. Packet loss rate

In this subsection, we discuss the packet loss of category $C1$'s measurements. The investigation serves for two purposes. The direct purpose is to gain a penetrating understanding of the Internet's IP-layer connectivity. So far, we have considered the IP-layer connectivity between a vantage point and a target to be equivalent to the success of reaching the target by the vantage point in its first round of direct probe. In fact, however, it is well-known that each packet is individually transferred, which means whether a packet to be dropped cannot be absolutely predicted by the successful arrivals of its former packets. Therefore, as the direct probe process actually consists of five probing packets, the percentage of category $C1$ itself is not a precise reference for evaluating the quality of the Internet services and applications that are sensitive to IP-layer packet loss. The other indirect purpose of this investigation is to check retrospectively to what extent the results in this study are dependent on our experiment implementation. For example, it will indicate a heavy dependence if most $C1$'s measurements have large packet loss rates, because it implies that had we designed the direct probe to use more probing packets, more vantage points would be able to reach more targets in their first round of direct probe and accordingly increase the category $C1$'s percentage in the final result.

Out of all the measurements in category $C1$, we find 95.05% had no packet loss at all in their direct probe processes, and 3.58%, 1.03%, 0.29%, and 0.05%, respectively, had one, two, three, and four lost packets. The results show that most vantage points and their reachable targets had reliable IP-layer paths to transfer packets. On the other hand, there are still around 5% IP-layer paths exhibiting high (over 20%) loss rates. Fig. 4 gives the detailed CDF plots of the percentages of category C1's measurements with different packet loss status among every vantage point. Over 95% of all vantage points had no packet loss occurring in their direct probe process towards more than 90% out of all targets. In contrast, a few other vantage points experienced packet loss as frequently as towards 60–70% of all targets, which was likely to be caused by some persistent congestion near the edge networks, through which these vantage points access to the Internet.

We also observe that the normalized amount of measurements that have different numbers of lost packets follow an exponential curve surprisingly well, as shown in Fig. 5. This exponential curve suggests that further increasing the number of probing packets in the direct probe would merely lead to negligibly small amount of additional targets to be classified into category $C1$. Therefore, it indicates that our experiment implementation of using five probing packets in the direct probe is already sufficient to make the study result considerably stable.

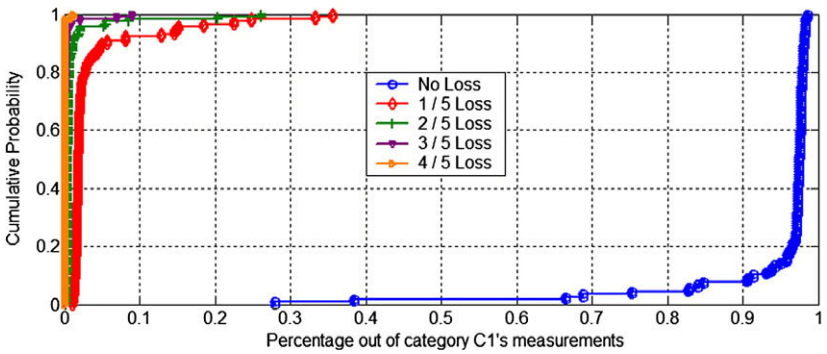

Fig. 4. The CDF plots of the percentages of category $C1$'s measurements with different packet loss status among typical vantage points.

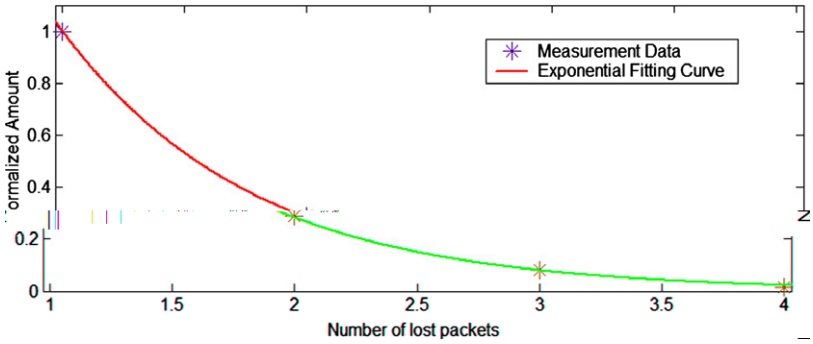

Fig. 5. The normalized amounts of measurements that have different numbers of lost packets fit an exponential curve with its function of $y = 3.55e^{-1.27x}$ precisely ($SSE = 1.06 \cdot 10^{-4}$, $R^2 = 0.9998$, $RMSE = 0.0073$).

5.1.2. Bursty characteristic of packet loss

Previous research has revealed that packets are lost with a highly bursty characteristic on the Internet [21,22]. In the following of this section, we revisit this characteristic based on our own measurements and prove its prevalent existence in a counter-evidence manner. In the presentation of our method, we use the term 'lost pattern' to refer to a certain permutation of binary symbols each of which indicates whether a corresponding packet is lost or not. For example, a lost pattern of '11000' means that only the first two out of all five probe packets are lost. Now assume that every packet is lost with a uniform probability and independently from each other. If this assumption is true, then given a specific lost rate, all possible lost patterns should have the same probability to appear. Based on this deduction, we can further use combinatories to figure out the ratio of the bursty pattern's appearing times relative to the number of all measurements. By bursty pattern, we mean those particular patterns with all the lost packets being contiguous. As an example, given all measurements having two out of five packets lost, the appearing times of four bursty patterns should be two fifths of the number of all measurements, because by combinatories there are totally ten possible loss patterns in this case. Table 4 gives the comparison between the ratios of bursty pattern's appearing times calculated based on above 'no-burst-characteristic' presumption and that obtained from our practical measurements. As can be seen, bursty patterns always appear notably more frequently in practice than it would be if packets were lost independently, which accordingly indicates the existence of the Internet IP-layer's bursty packet-loss characteristic.

5.2. Delay variation

In this subsection, based on category $C1$'s partial measurements that had no packet loss at all, we inspect into the E2E delay variation on the Internet. So far, a lot of research has been devoted into finding suitable models to characterize the Internet's E2E delay behavior, from elaborate models as complicated as using system identification and time series analysis [23,24], to practical models as simple as just predicting with the minimum or median of a few most recent measurements [25]. In many applications, the purpose of predicting the Internet's E2E delay is to design a system working more stably and more efficiently. We expect our in to gain useful insight in the sense that if the delay variation is small in most cases, application developers can prefer simple and practical models with confidence that they are able to predict E2E delay considerably accurately. Besides, our study also distinguishes itself from existing experimental studies by achieving topological and geographic diversity among the measured IP-layer paths, instead of sampling just a few paths repeatedly over time. Obviously, this diversity is necessary to a thorough understanding of the Internet's delay behavior.

Table 4
Comparison between the ratios of bursty pattern's appearing times calculated based on 'no-burst-characteristic' presumption and that observed in practical measurements.

Loss rate	2/5 (%)	3/5 (%)	4/5 (%)
No-burst-loss	40	30	40
In practice	42.13	35.17	57.28

The E2E delay on the Internet's IP-layer mainly consists of three components, the processing time of terminals, the processing and queuing time of intermediate network devises, and the propagation time of the radio or light signal transferring through the media. Usually, the dominant component is the propagation time, whose best estimate is often considered to be the minimum of a given series of measurements. As a more comprehensive prediction of the E2E delay, the median of these measurements instead of their mean is more appropriate, because it is clear that the Internet's E2E delay follows a single-side-heavy-tail distribution that makes the mean easily biased by a few extremely large 'flying points'. The delay variation is usually caused by either the change of route or the dynamic work load of relevant equipments.

Intending to provide rich reference to a variety of Internet services and applications that have different sensitivities and requirements on delay variation behaviors, we study four different metrics to measure delay variation, namely the standard deviation, difference between maximum and minimum, difference between mean and median, and difference between median and minimum. Fig. 6 illustrates the CDF plots of the absolute delay variation. As can be seen, with any one of the four studied metrics, a large percent of delay variations are no more than 10 ms. On the other hand, however, there are still a considerable proportion of delay variations that are surprisingly large. For example, we find a series of measurements from a Netherlands vantage point'planetlab1.ewi.tudelft.nl' to a South Africa target '196.32.164.1' was '242, 8560, 7560, 6565, 5570'. According to the distinct increase between the first and the following four measurements, there seemed to be a prominent route change or persistently critical congestion occurring at that time, but surprisingly no packet had been lost during the congestion period. This situation could lead to some unexpected consequence on certain degradation-based delay prediction models, in which after the congestion had been eliminated, the bias impact of many extremely large delay measurements would take a long time to regress.

5.3. Overlay routing implication

According to Table 3, around 25–50% of the unreachability could be recovered by delegating the probe to another proper van-

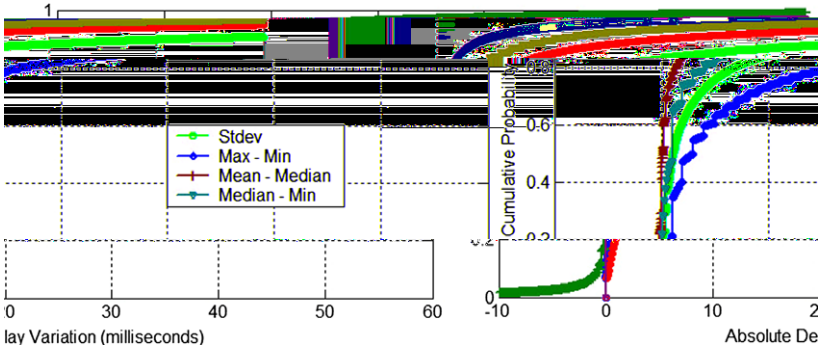

Fig. 6. The CDF plots of delay variation in terms of four different metrics.

tage point to detour the IP-layer direct unreachability between the original vantage point and the target. In this subsection, we investigate to what extent our implementation of the delegation process influences this result, and what is its implication to the understanding of one-hop overlay routing techniques.

In the delegation process in our experiment, at most five one-hop overlay paths were attempted to reach the target. Among the measurements in which the IP-layer outage were successfully detoured in the delegation process, 79.18% attempted merely one other randomly chosen vantage point. Fig. 7 illustrates the normalized amount of measurements in which the IP-layer outage was successfully recovered by attempting different numbers of overlay paths. We note the data well fit an exponential curve. According to this curve, increasing the number of attempted overlay paths in the delegation process would only detour negligibly additional amount of IP-layer outage between the original vantage point and the target. Therefore, we argue that our implementation of the delegation process neither sways the related measurement results, nor obstructs their use for reference to understand the effect of one-hop overlay routing. Our result is consistent with Gummadi et al.'s previous research in [26], where they also found that randomly selecting four intermediaries is the best tradeoff between recovery effort and the success of one-hop overlay routing technique. Given their conclusion was based on contiguous measurements among a small set of end hosts, our experiment further proves it to be a common sense prevalently existing on the Internet.

5.4. Domain effect

Considering that the Internet's IP routing service is running on a domain-based hierarchy, in this subsection we investigate whether and to what extent the domain effect influences the Internet's IP-layer connectivity. In particular, we first compare the statistical characteristics of the measurements with their vantage points and targets located in the same domain (AS/PoP) to the statistical characteristics of general measurements as we have studied previously. Then, we study whether or not the Internet's IP-layer connectivity to the targets in the same domain is correlated with each other.

5.4.1. Domain effect on statistical characteristics

We note two main reasons that may cause domain effects impacting the Internet's IP-layer connectivity. For one thing, the intra-domain topology is often much more richly connected than the inter-domain topology. This is because the physical links inside the same domain are mostly short-range and their deployment does not necessitate business contracts between different ISPs. For another, different requirements make intra- and inter-domain routing protocols choose vastly different routing styles. While intra-domain routing merely needs to care about robustness and performance, inter-domain routing faces a formidable combination of algorithmic and policy challenges due to the economical reasons for supporting ISPs to flexibly implement their private routing policies. As a result, intra-domain routing protocols such as OSPF and IS–IS often make use of the link-state style of routing, which not only can find the optimal path in terms of a specific metric, but also has many other advantages including fast convergence, incurring low churn, and easy fault diagnosis. However, link-state routing style is unsuitable to support policy-based routing, because it has to reveal every network activity to all participants and violate privacy norms of policies. Due to this reason, the current inter-domain routing protocol BGP chooses to use path-vector routing style, which enables complex policies and suppresses loops by advertising full path information to the destination. It is known that BGP suffers from significant route instabilities, route oscillation and long convergence time [27]. Consequently, we have reasons to conjecture that the Internet's IP-layer connectivity within

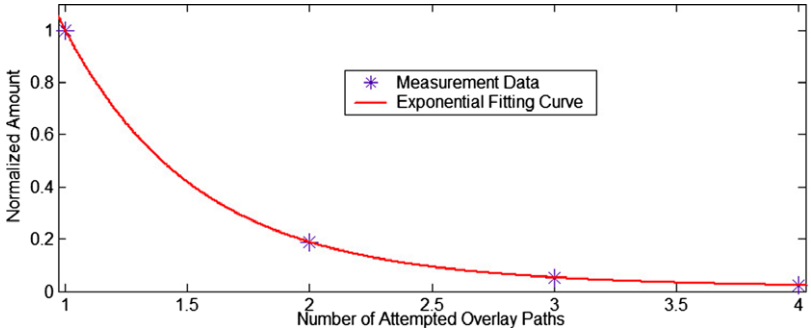

Fig. 7. The normalized numbers of measurements that successfully detoured IP-layer outage by attempting different numbers of one-hop overlay paths. They fit an exponential curve with its function of $y = 6.2e^{-1.92x} + 0.15e^{-0.51x}$ precisely ($SSE = 1.78 \cdot 10^{-32}$, $R^2 = 1$).

the same domain should outperform that between different domains. In regard to the study in this paper, the domain effect can be reflected by the fact that a vantage point can reach larger percent of targets that located in the same domain than those in different domains, and that the vantage point has smaller packet loss rate and delay variation to the targets in the same domain.

Table 5 shows the percentage of each category among different sets of targets. Clearly as it is, the percentage of the directly reachable category $C1$ in the sets where the targets are in the same AS/PoP as the vantage point is larger than that in the set of all targets. Due to the limit of space, we only give the results of original measurements, but we have examined that the conclusion also remains true with the adjustment methods proposed in Section 4.2. Table 6 details the constitution of the measurements in category $C1$ according to different packet loss situations. Also as expected, it indicates that larger percent of measurements towards the targets in the same AS/PoP have no or less packet loss relative to the measurements towards all targets. Similar comparative result can also be observed from the CDF plots of delay variations of category $C1$'s partial measurements that have no packet loss, as shown in Fig. 8. In summary, these results validate the existence of domain effects in terms of reachability, packet loss and delay variation.

Another notable observation is that PoP-level's domain effect is much more outstanding than that of AS-level's. We find that one of the main reasons is the existence of some extremely large ASes, each of which distributes its PoPs all over the world. For example, we observe that AS 1239 which belongs to Sprint ISP has PoPs not only in U.S., but also in European and Asian Pacific regions. As a result, two hosts located in the same one of these ASes may still have a long geographic distance from each other, and thus the IP path between them may consist of a great number of routers. According to our measurements, the RTT minimum is mostly smaller than 20 ms in the same PoP, 100 ms in the same AS, but ranges from dozens of to hundreds of milliseconds if including inter-domain measurements.

Table 5
The domain effect on connectivity makes the reachable percent of targets in the same domain larger than that of all targets.

Category	$C1$ (%)	$C2$ (%)	$C3$ (%)	$C4$ (%)	$C5$ (%)
All targets	95.01	0.59	0.33	4.05	0.02
Targets in same AS	96.01	1.43	0.32	2.23	0.01
Targets in same PoP	98.20	1.09	0.00	0.71	0.00

Table 6
The domain effect makes the packet loss rate towards targets in the same domain smaller than that towards all targets.

Loss rate	*No Loss* (%)	1/5 (%)	2/5 (%)	3/5 (%)	4/5 (%)
All targets	95.05	3.58	1.03	0.29	0.05
Targets in same AS	95.43	3.25	0.92	0.36	0.04
Targets in same PoP	99.44	0.45	0.11	0.00	0.00

5.4.2. Domain effect on connectivity correlation

We next study the domain effect on the correlation of the Internet's IP-layer connectivity by verifying whether the targets in the same domain are likely to be reachable by the same subset of vantage points. To this end, for each target t we define a connectivity vector $V_t = [R_{ti}](1 \leqslant i \leqslant 118)$ according to the measurements towards the target t from every vantage point i. R_{ti} equals to 1 if the measurement from vantage point i towards target t belongs to category $C1$, i.e. vantage point i can successfully reach target t in its first round of direct probe; otherwise, R_{ti} equals to 0. Afterwards, given a set of targets $t_1, t_2, \ldots, t_n$, we use the linear dependence over their corresponding connectivity vectors $V_{t_1}, V_{t_2}, \ldots, V_{t_n}$ to denote the correlation between their IP-layer connectivity. Specifically, we form a linear space with vectors $V_{t_1}, V_{t_2}, \ldots, V_{t_n}$, and define the ratio of the linear space's rank relative to the number of vectors, that is n, as the independence factor of these vectors to quantitatively measure their linear dependence. Under such a formalization framework, the domain effect on connectivity correlation can be interpreted as that the targets in the same domain have statistically smaller independence factor than those targets in different domains.

Fig. 9 illustrates the CDF plots of the independence factors of different sets of targets. It is important to note that these statistical results are not based on all targets, but those after some necessary filtration. First, we exclude the targets that are reachable by every vantage point, since their connectivity vectors are all the same. Second, after clustering the remained targets into a number of domains, we further filter out the extremely small domains (including less than 10 targets) to reduce the interference of accidental causes, such as fault inference of a target's domain and occasional unreachability due to bursty packet loss. Ultimately, the AS-level statistics are based on totally 60135 targets scattered in 1411 different ASes, and the PoP-level statistics are based on 20963 targets in 905 different PoPs. Moreover, if a large domain include more than 118 (the length of the connectivity vector) targets, we repeatedly select 118 random targets from the domain multiple times in proportion to the number of targets the domain includes. Every time we calculate the independence factor of a number of targets in the same domain, and then we randomly picked out the same number of targets from the whole target pool and calculate their independence factor for contrast. As shown by Fig. 9, the independence factors of targets in the same AS/PoP are statistically smaller than their counterparts, indicating that the IP-layer connectivity in the same domain is evidently correlated. As a practical implication, it suggests that domain diversity is significant for increasing geo-

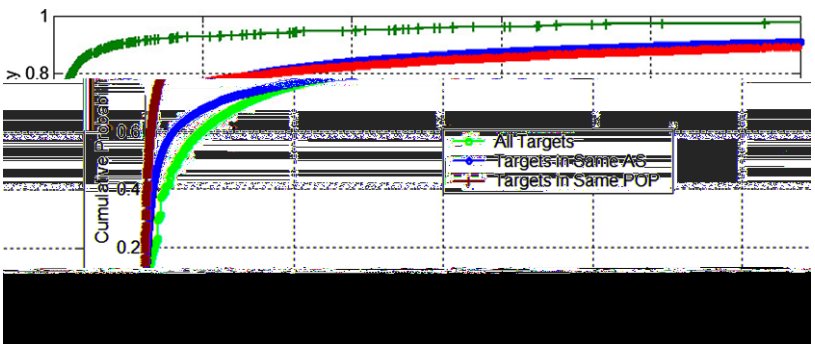

Fig. 8. The domain effect makes the delay variation towards targets in the same domain statistically smaller than that towards all targets.

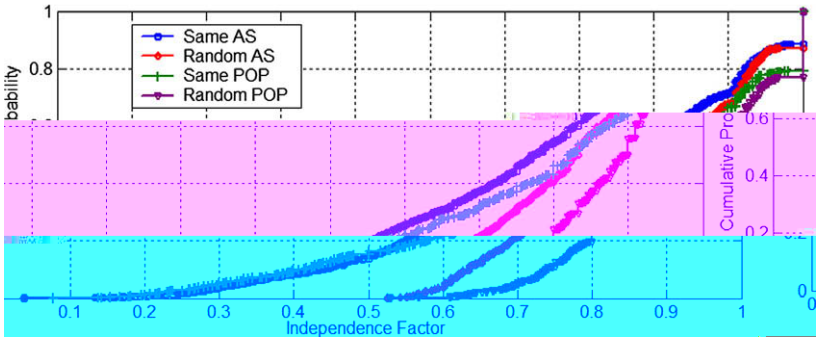

Fig. 9. The domain effect makes the targets in the same domain have statistically smaller independence factors than the randomly chosen targets.

graphic reliability of Internet applications such as the Web content provisioning.

5.5. Geographic distance effect

To study the impact of geographic distance between the vantage point and the target on their IP-layer connectivity, we take the Earth as a globe and measure the geographic distance with radians of the arc passing through the vantage point's and the target's inferred locations. In total, we successfully inferred the latitudes and longitudes of 45933 targets, and the measurements towards each of these target from every vantage point accordingly compose the dataset for this study.

We compare the statistical geographic distances between the vantage point and target in measurements of different categories, with different packet loss and with different delay variation, respectively. The results show that the geographic distance indeed has a little correlation to the Internet's IP-layer connectivity, in the sense that the sets of measurements with better reachability and having smaller packet loss and delay variation all statistically have shorter geographic distances. However, such correlation is too weak to take the geographic distance as a competent heuristic for estimating the Internet's IP-layer connectivity.

We also reconfirm the prevalent existence of a coarse-grain correlation between the delay and geographic distance. The correlation was first pointed out in [28] based on an experiment using vantage points in 14 locations in U.S. to probe 265 Web servers that spread across university campuses in 44 states. While our study does not essentially extend the conclusion, it supplements the understanding of this issue in at least two aspects. First, as the authors indicated that their experiment had been limited to a particular domestic network environment that merely consisted of U.S. university sites. Because most sites were well connected to the high-speed Internet2 backbones, they were far from representative for the Internet's heterogeneity. Second, the Internet has grown rapidly in recent years, and has been more than doubled in terms of the number of IP prefixes since their experiment was carried out. It is uncertain whether and how the Internet's scale expansion would have changed its conventional behaviors. Given these indefinability, our result reveals that the positive correlation between the delay and geographic distance is still a common sense prevalently existing on the current Internet. But it is also observed that the correlation becomes weaker and weaker as the network delay increases, especially after the RTT is larger than 150 ms.

6. Special cases with surprisingly high unreachability

As indicated in Section 4.1, while 118 out of 124 vantage points could reach more than 90% of all targets in their first round of direct probe, there are still six other vantage points having strikingly high IP-layer unreachability. In this section, we inspect these special cases based on both analysis and empirical study.

6.1. Conjecture and analysis

In Table 7, the results related to the first experiment show the original constitution of the measurements from each of the 6 special vantage points. Comparing it with Table 3, we can observe that the 6 special vantage points mainly exhibit two characteristics.

The first characteristic is typified by the first 2 vantage points, which have a large proportion of measurements in both category $C2$ and $C3$. The considerably large percentage of category $C3$ indicates that the IP-layer unreachability from these vantage points towards many targets was successfully recovered in the confirmation process. Given this, it seems that most of such ephemeral IP-layer unreachability was caused by critical network congestion incidents which had resulted in vastly high packet loss rate. Convincingly, we indeed find that the average loss rates of the category $C1$'s measurements from these two vantage points are, respectively, the third and seventh highest among all 124 vantage points.

The second characteristic is typified by the rest 4 vantage points, each of which has a large proportion of measurements in category $C2$. It means all these vantage points could not reach a large portion of targets in their first round direct probe process, but could in their delegation process. Although it is possible that

Table 7
The constitution of the measurements from the 6 special vantage points in two experiments apart one month (the values are interpreted as *first-experiment-result%/second-experiment-result%*). The increase of category $C4$'s percentage in the second experiment is because some targets might become absent/offline after the first experiment.

Vantage point	$C1$	$C2$	$C3$	$C4$	$C5$
planetlab1.engr.uconn.edu	83.98/80.76	4.74/5.06	6.87/7.78	4.38/6.35	0.02/0.05
planetlab1.cse.msu.edu	81.29/93.30	9.15/0.25	5.73/0.19	3.80/6.24	0.03/0.02
planetlab1.cs.unibo.it	68.01/66.70	27.87/26.00	0.31/0.43	3.80/6.84	0.01/0.03
planetlab2.win.trlabs.ca	66.69/63.25	28.88/29.62	0.11/0.09	4.30/7.01	0.02/0.02
pli2-pa-3.hpl.hp.com	45.81/40.64	50.15/52.40	0.09/0.06	3.94/6.89	0.01/0.01
pl1.unm.edu	0.00/0.00	95.96/93.97	0.00/0.00	4.04/6.03	0.00/0.00

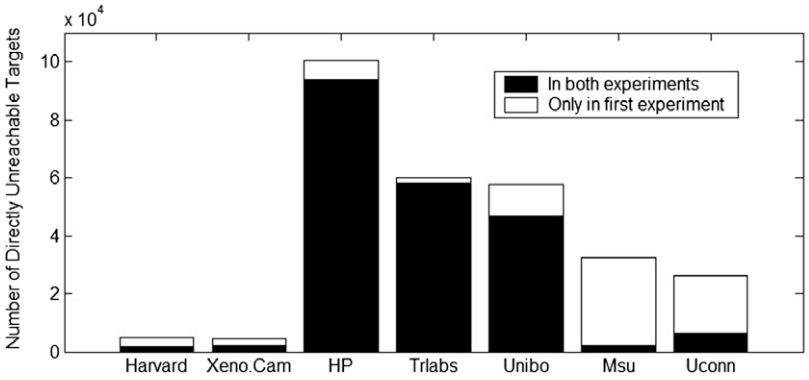

Fig. 10. The number of directly unreachable targets. As the cases of the typical vantage points are similar to each other, we only present two of them for comparison purpose. For clarity of the figure, we does not show the extremely large numbers of the particular vantage point *pl1.unm.edu*.

such a high IP-layer direct unreachability was due to abnormally accidental failures, more likely it was caused by intentional security or routing policies.

6.2. Verification and empirical study

To verify whether or not the above analysis is correct, we did our experiment again on these 6 special vantage points and 6 other randomly chosen typical vantage points one month later in February, 2008.

We first infer the causes of IP-layer unreachability by investigating the variation of the directly unreachable targets between the two experiments. Intuitively, given a vantage point, if its directly unreachable targets in the first experiment became directly reachable in the second experiment, then it is likely that the previous unreachability was due to accidental reasons such as the routing misbehavior and critical congestion; otherwise, had some targets been directly unreachable in both experiments, the unreachability was more likely due to intentional insulation reasons such as the security or routing policies set by network operators.

Fig. 10 shows the composition of directly unreachable targets from each vantage point. As it shows, for both vantage points *planetlab1.engr.uconn.edu* and *planetlab1.cse.msu.edu*, a majority of the directly unreachable targets in the first experiment became directly reachable in the second experiment. It implies that the high IP-layer unreachability from these two vantage points in the first experiment was unlikely caused by intentional insulation. Moreover, as shown in Table 7, the category *C*1's percentage of *planetlab1.cse.msu.edu* had notable increase in the second experiment, becoming comparable to those of typical vantage points, but that of 'planetlab1.engr.uconn.edu' did not. Therefor, the unreachability from *planetlab1.cse.msu.edu* was likely due to critical congestion happened near the vantage point's access networks, while that from *planetlab1.engr.uconn.edu* was mainly due to the inferior performance of access networks.

In contrast, the other 4 special vantage points all had a majority of directly unreachable targets in the first experiment also directly unreachable in the second experiment. Therefore, most unreachability from these vantage points was likely due to routing or security policies that had been intentionally set by network operators. Next, we present some interesting observations based on our manual investigation.

At the first glance, *pl1.unm.edu* seems the most special in Table 7. It failed to reach all except two targets through its direct IP-layer probe, but more than 95% of these directly unreachable targets were successfully reached by using another proper vantage point in the delegation process. We find that this particular situation was caused by some subtle security policies that forced the border routers of this vantage point's domain to drop either the egress ICMP ECHO_REQUEST or ingress ICMP ECHO_RESPONSE packets. Our conclusion is based on the following experimental observations. First, the only two targets that were directly reachable by *pl1.unm.edu* were both located in the same AS as *pl1.unm.edu*. Moreover, if we let *pl1.unm.ed* `ping` each hop's IP address in the output of `traceroute` to a target, all the time there was a boundary hop before which every hop was reachable and after which every hop was not. The boundary hop was always the first IP address belonging to another AS different from the one where *pl1.unm.edu* located in. We also did the experiment reversely from several other vantage points towards *pl1.unm.edu*, and similar boundary hops existed. On the other hand, however, although *pl1.unm.edu* could not reach other vantage points in different domains through `ping`, we found that it could successfully communicate with them using UDP and TCP protocols. Recall that the direct probe process was based on ping and ICMP protocol, while the delegation process was carried out through SSH over TCP protocol; it explains why more than 95% of the directly unreachable targets could be reached in the delegation process.

Another interesting observation is when executing `traceroute` on vantage point *pl1.unm.edu*, it never reached the final destination successfully. To give an example, suppose there was a `traceroute` output containing the following sub-route *pl1.unm.eduAB**. Although it indicated that *B* was reachable, if we next executed `traceroute` with *B* as the target, we would get *pl1.unm.eduA** instead of the expected result *pl1.unm.eduAB*. This quaint phenomenon was likely caused by a security policy that conducted the relevant routers to drop 'ICMP Port Unreachable' packets but not to drop 'ICMP Time-To-Live Exceeded' packets, because `traceroute` triggers the former type of messages only at the final destination, but the latter type on all the other hops. We suppose this security policy was probably set up to avoid port scanning attacks. While this security policy has not direct impact on the study results in this paper since we actually used `ping` as the probe method to check IP-layer connectivity, we still report this limit of `traceroute` to make researchers who intend to carry out similar studies be aware of the possible misleading effect of this particular situation.

Finally, we find that a great number of targets being directly unreachable from *pli2-pa-3.hpl.hp.com* were due to an unusual routing black-hole. It was observed that many targets reachable through both `ping` and `traceroute` by other vantage points could not be reached by *pli2-pa-3.hpl.hp.com*. Whenever executing `traceroute` to one of these targets from *pli2-pa-3.hpl.hp.com*, it always stopped since the fifth hop and could get no response from the following hops except timeout. The last traceable router inter-

face along the routes to these directly unreachable targets was *cenic-rtr.Stanford.EDU*, where there seemed to be a routing sink and the router did not know to which next hop it should forward the packets. As this problem existed in both experiments over one month apart, it was unlikely caused by physical link failures or routing pathologic behaviors. Most probably, it was due to the lack of proper routing information to these directly unreachable targets caused by certain improper configuration of BGP routers.

7. Routing issues impacting IP-layer unreachability

In this section, we analyze how our measurement results are correlated to typical routing issues.

7.1. Inter-domain routing policies

The current Internet's routing infrastructure consists of a two-layer hierarchy. On the first layer, BGP is used as the inter-domain protocol that exchanges information between different ASes to announce, update, and withdraw AS paths to reach publicly routable IP prefixes. On the second layer, intra-domain protocols such as OSPF and IS–IS are used to establish and maintain the optimal route that can pass a packet through a series of routers in the present AS to a router belonging to the next hop AS.

As ASes are often separately operated by different ISPs, the inter-domain protocol BGP is designed in a policy-based style to fit AS relationships that are determined by commercial agreements between relevant ISPs. In general, the AS relationship has three different types [29]: *provider-to-customer*, *peer-to-peer*, and *sibling-to-sibling*. In the first case, a customer AS pays a provider AS to connect to the Internet and transit traffic to and from other ASes. With *peer-to-peer* relationship, two ASes agree to exchange their own traffic and the traffic from their respective customer ASes free of charge. The *sibling-to-sibling* relationship is set between ASes that belong to the same ISP and fully cooperate on sharing routing information and tuning traffic.

Notably, based on these AS relationships, there is a widely adopted routing policy named 'valley-free' [30] that in theory can lead to IP-layer unreachability between two ASes. Under the 'valley-free' routing policy, customer ASes do not transit traffic from one provider AS to another, and peer ASes do not exchange traffic coming from other peer ASes. As a result, if two ASes neither directly connect to each other, nor, respectively, inherent (no necessary to be directly) from two provider ASes that can exchange traffic on behalf of their customer ASes, then these two ASes will suffer from IP-layer unreachability. In regard to our experiment results, if a vantage point and a target were, respectively, located in such two ASes, the vantage point would not be able to reach the target in both the first direct probe and confirmation processes; in the delegation process, as the attempted delegate vantage points were distributed in a diversity of ASes, it was likely that the target could be successfully reached. Therefore, most of this type of IP-layer unreachablity would be classified into category *C*2, and negligibly percent might be classified into category *C*4.

Although possibly existing in theory, the above AS pairs having no 'valley-free'AS paths should seldom or just temporarily appear in practice, because most customer ASes are supposed to connect to at least one of the full-mesh-connected Tier-1 ASes. Instead, routing dynamics, as discussed in the next subsection, are often supposed more responsible for IP-layer unreachability.

7.2. Routing dynamics

Routing dynamics are mainly triggered by two types of events: network congestion and physical topology change. When severe network congestion happens, the overloaded router has to drop many packets that arrive in a burst. The performance degradation or even temporary interrupt of the corresponding link may trigger the upstream router switching to another recalculated route. If such congestion happened on the route between a vantage point and a target in our experiment, both the congestion itself and the resulted route switch could cause IP-layer unreachability. However, considering that many traffic load-balancing and fast reroute schemes [31] have been adopted in the Internet's routing infrastructure, we believe that this type of IP-layer unreachability is most likely to be ephemeral and therefore mostly classified into category *C*3 (if congestion was successfully detoured) or *C*5.

Physical topology change can happen due to various reasons, such as establishment or abolishment of commercial agreements between ISPs, intentionally adding/removing routers/links to maintain or upgrade the networks, and most commonly the accidental failures of network devices. After the physical topology changes, the involved intra-domain and inter-domain routing protocols will try their best to recalculate another suitable route to transmit relevant traffic. If no new route could be found, it would lead to persistent IP-layer unreachability, which should have been classified into category *C*2 (if the route failure was successfully detoured) or *C*5 in our experiment. Even if there existed such a suitable new route, the IP-layer reachability also could be interrupted for a while before the routing state converged again, because it took time to discover the link failure and propagate routing information across the networks. In general, if the routing dynamics are limited in intra-domain scope, the IP-layer unreachability can be recovered in a relatively short term, as intra-domain routing protocols are mostly based on link-state algorithms in which every router maintains a global topology graph. In this case, the observed IP-layer unreachability in our experiments was likely to be classified into category *C*3 or *C*5.

On the other hand, however, if the routing dynamics involve changing the AS path, the inter-domain routing protocol BGP may take a long time to converge. To alleviate the router's processing load, BGP uses a minimum router advertisement interval (MRAI) timer to determine the minimum amount of time between sending the same neighbor two continuous routing updates to a particular destination prefix. While effective on reducing the number of updates triggered by each routing event, MRAI timer can also defer routing convergence, and if implemented improperly, it can even lead to long-lived routing black holes. Previous study has revealed that various inter-domain routing issues including routing policies, iBGP configurations, MRAI timer values, and failure locations, can all have significant impact on the Internet's routing failures [32]. In regard to our experiment, a majority of the IP-layer unreachability caused by inter-domain routing dynamics was most likely classified into category *C*2, and a small portion into *C*4.

To gain a general understanding of routing dynamics, we performed another experiment on PlanetLab to measure route changes. Specifically, we deployed a set of scripts on each of the 278 PlanetLab nodes, as selected in Section 2.1. The scripts first conducted each node to initialize a list of the other PlanetLab nodes that could be successfully reached with `traceroute`, and then used `traceroute` to monitor every route from the current node to each of the reachable nodes in the initialization stage. The time interval between two consecutive measurements of the same route was a constant minimum (5 min) plus an independent and exponentially distributed random variable; roughly, the average time interval was one hour. In particular, if the destination was unreachable at some time, the next measurement would be launched after the minimum time interval. In each measurement, the scripts executed `traceroute` with the following parameters: for each hop, two probe packets were sent with a minimum interval of 500 ms; each probe packet's maximum time-to-live (TTL)

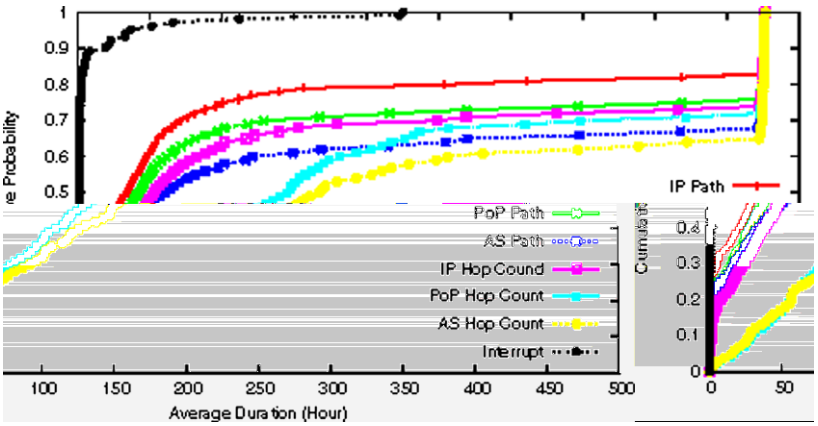

Fig. 11. The CDF plots of the route's persistent and interrupt duration averaged by each pair of PlanetLab nodes.

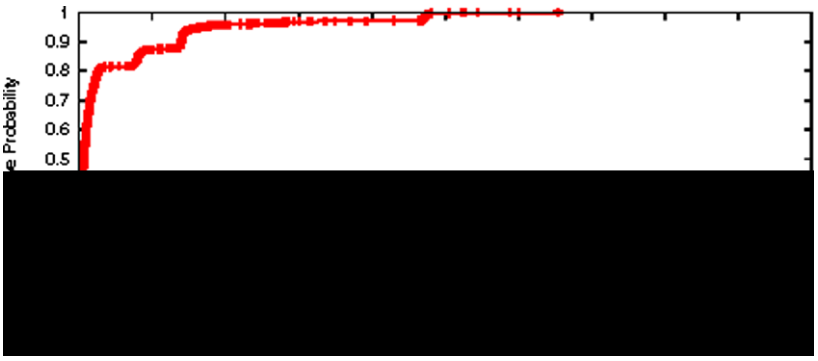

Fig. 12. The CDF plot of the ratio of every observed pair's routing outage time to the whole monitoring time.

was 50. By running the experiment for around three weeks, we finally collected the monitor results of the routes between 11368 different pairs of PlanetLab nodes.

Fig. 11 illustrates the CDF plots of the average persistent duration of the monitored routes. Given two consecutive `traceroute` measurement results between the same pair of nodes, we determine whether the route had changed with an optimistic strategy to treat possible anonymous hops in the `traceroute` outputs. Specifically, only if at least one hop in the hinder measurement is explicitly different from its counterpart in the previous measurement, will we judge that the route ever changed in the middle of the two measurement's time. Moreover, as we cannot determine how long the route had persisted before the first route change or would persist after the last route change, we exclude out these two types of samples from the calculation of each route's average persistent duration. Every plot except the 'Interrupt' one in Fig. 11 is comparable to each other, because they are all based on the statistics of 9413 pairs of PlanetLab nodes, which is the intersection of all scenarios. As can be seen, the average route persistent duration can vary vastly between different pairs. While a few pairs have their average route persistent duration as short as several minutes, around 20% pairs have theirs as long as tens of days. Indeed, we find many pairs whose route had never changed during the whole experiment (around three weeks). As another reasonable observation, the routing properties of coarser grain are more stable and persist longer than those of finer grain, such as the AS/PoP path compared to IP path and the route's hop count compared to its specific hop sequence.

Among the total 9413 pairs, there are 2462 pairs ever suffering from route outage. The last plot in Fig. 11 shows the statistics of the average IP-layer interrupt duration between every pair. More than 50% pairs have their average interrupt duration shorter than 1 h, 90% shorter than 10 h; on the other hand, however, there are also a few pairs having extremely long interrupt duration, even more than a week. Fig. 12 shows the ratio of every observed pair's routing outage time to the whole monitoring time during our experiment. The average failure ratio is 4.02% for the pairs ever suffering from routing outage, and is 1.05% for all the 9413 pairs.

One may note that compared to the large number of targets studied in Section 3.1, the route monitoring experiment was unscalable and only able to collect route dynamics to a much smaller number of targets. However, due to the diverse distribution of PlanetLab nodes, we believe the above statistics are still representative. Moreover, there have been many other measurement studies correlating the routing instability and E2E performance [32,27,33–35]. All these indicate that routing dynamics can lead to IP-layer unreachability lasting for a wide variation of time.

8. Concluding remarks

This paper investigates the situation of the current Internet's IP-layer connectivity. Given the tremendous scale, complexity, and heterogeneity of the Internet, we presented our methodology and experiment design on how to collect representative measurements to study the IP-layer reachability.

Quantitative study on around two hundred million measurements shows that the Internet's average IP-layer connectivity is around 95–98%. Specifically, if running an Internet service on one of the typical 118 vantage points played by PlanetLab nodes that are scattered all over the world, the likelihood of being able to successfully reach the service on the IP-layer from different access networks in different locations ranges from 90.3% to 95.9%.

Among the directly reachable measurements, 95.05% have no packet loss at all, and the others prevalently exhibit a bursty characteristic of packet loss. The delay variation of the directly reachable measurements having no packet loss exhibits a long-tail distribution. While the delay variation of over 90% measurements is no more than 10 ms, there are still a few measurements with extremely large delay variation. Around 25–50% of the IP-layer unreachability can be successfully detoured by using another proper vantage point, and thus the outage is likely caused by backbone problems. Moreover, notable domain effect is observed on the Internet's IP-layer connectivity. For one thing, the measurements towards the targets located in the same domain as the vantage point have larger proportion directly reachable and statistically with smaller packet loss rate and delay variation. For another, the connectivity to targets located in the same domain is more correlated with each other than that to targets in different domains. Besides, the coarse-grain correlation between delay and geographic distance is reconfirmed to prevalently exist on the current Internet.

Finally, the paper investigates main causes that can lead to IP-layer unreachability, including intentional insulation policies revealed in the empirical study and accidental interrupt related to the Internet's routing issues.

References

[1] D.G. Andersen, Improving end-to-end availability using overlay networks, Ph.D. Thesis, 2005.
[2] V. Paxson, End-to-end routing behavior in the internet, IEEE/ACM Transactions on Networking 5 (5) (1997) 601–615.
[3] N. Feamster, D. Andersen, H. Balakrishnan, M.F. Kaashoek, Measuring the effects of internet path faults on reactive routing, in: Proceedings of ACM SIGMETRICS'03, 2003.
[4] W. Jiang, H. Schulzrinne, Assessment of voip service availability in the current internet, in: Proceedings of PAM'03, 2003.
[5] Comon. Available from: <http://comon.cs.princeton.edu/>.
[6] J. Pang, A. Akella, A. Shaikh, E. Krishnamurthy, S. Seshan, On the responsiveness of DNS-based network control, in: Proceedings of ACM IMC'04, 2004.
[7] Ipv4. Available from: <http://www.potaroo.net/tools/ipv4/>.
[8] H.V. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, A. Venkataramani, iPlane: an information plane for distributed services, in: Proceedings of USENIX OSDI'06, 2006.
[9] A. Broido, kc claffy, Analysis of routviews BGP data: policy atoms, in: Proceedings of Network Resource Data Management Workshop, 2001.
[10] F. Baker, Requirements for ip version 4 routers, rFC 1812 (June 1995).
[11] Internet registries. Available from: <http://www.iana.org/ipaddress/ipaddresses.htm>.
[12] Routeviews. Available from: <http://www.routeviews.org/>.
[13] Ripe. Available from: <http://www.ripe.net/>.
[14] N. Spring, R. Mahajan, D. Wetherall, T. Anderson, Measuring ISP topologies with rocketfuel, IEEE/ACM Transactions on Networking 12 (1) (2004) 2–16.
[15] Sarangworld. Available from: <http://sarangworld.com/TRACEROUTE/patterns.php3>.
[16] Geoworldmap. Available from: <http://www.geobytes.com/>.
[17] BGP analysis reports. Available from: <http://bgp.potaroo.net/>.
[18] C. Labovitz, A. Ahuja, F. Jahanian, Experimental study of internet stability and wide-area backbone failures, in: Proceedings of FTCS'99, 1999.
[19] M. Dahlin, B. Chandra, L. Gao, A. Nayate, End-to-end wan service availability, IEEE/ACM Transactions on Networking 11 (2) (2003) 300–313.
[20] G. Huston, Analyzing the internet BGP routing table, The Internet Protocol Journal 4 (1) (2001). Available from: <http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_4-1/bgp_routing_table.html>.
[21] M.S. Borella, D. Swider, S. Uludag, G.B. Brewster, Internet packet loss: Measurement and implications for end-to-end qos, in: Proceedings of ICPP'98, 1998.
[22] W. Jiang, H. Schulzrinne, Modeling of packet loss and delay and their effect on real-time multimedia service quality, in: Proceedings of NOSSDAV'00, 2000.
[23] H. Ohsaki, M. Murata, H. Miyahara, Modeling end-to-end packet delay dynamics of the internet using system identification, in: Proceedings of the International Teletraffic Congress 17, 2001.
[24] M. Yang, A. Bashi, J. Ru, X. Rong, L.H. Chen, S. Rao, Predicting internet end-to-end delay: a statistical study, The Annual Review of Communications 58 (2005) 665C677.
[25] L. Tang, H. Zhang, J. Li, Y. Li, End-to-end delay behavior in the internet, in: Proceedings of MASCOTS'06, 2006.
[26] K.P. Gummadi, H.V. Madhyastha, S.D. Gribble, H.M. Levy, D. Wetherall, Improving the reliability of internet paths with one-hop source routing, in: Proceedings of USENIX OSDI'04, 2004.
[27] C. Labovitz, A. Ahuja, A. Bose, F. Jahanian, Delayed internet routing convergence, IEEE/ACM Transactions on Networking 9 (9) (2001) 293–306.
[28] V.N. Padmanabhan, L. Subramanian, An investigation of geographic mapping techniques for internet hosts, in: Proceedings of ACM SIGCOMM'01, 2001.
[29] X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, kc claffy, G. Riley, As relationships: inference and validation, ACM SIGCOMM Computer Communication Review 37 (1) (2007) 29–40.
[30] Z.M. Mao, L. Qiu, J. Wang, Y. Zhang, On as-level path inference, in: Proceedings of ACM SIGMETRICS'05, 2005.
[31] E. Osborne, A. Simha, Traffic Engineering with MPLS, Cisco Press, 2002.
[32] F. Wang, Z.M. Mao, J. Wang, L. Gao, R. Bush, A measurement study on the impact of routing events on end-to-end internet path performance, in: Proceedings of ACM SIGCOMM'06, 2006.
[33] M. Roughan, T. Griffin, M. Mao, A. Greenberg, B. Freeman, Combining routing and traffic data for detection of ip forwarding anomalies, in: Proceedings of ACM SIGCOMM NeTs Workshop, 2004.
[34] N. Feamster, D.G. Andersen, H. Balakrishnan, M.F. Kaashoek, Measuring the effects of internet path faults on reactive routing, in: Proceedings of ACM SIGMETRICS'03, 2003.
[35] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, C. Diot, Characterization of failures in an ip backbone, in: Proceedings of IEEE INFOCOM'04, 2004.