A Measurement Study on Potential Inter-Domain Routing Diversity

Chengchen Hu, *Member, IEEE,* Kai Chen, Yan Chen, *Member, IEEE,* Bin Liu, *Senior Member, IEEE,*
Athanasios V. Vasilakos, *Senior Member, IEEE*

***Abstract*—In response to Internet emergencies, Internet resiliency is investigated directly through an autonomous system (AS) level graph inferred from policy-compliant BGP paths or/and traceroute paths. Due to policy-driven inter-domain routing, the physical connectivity does not necessarily imply network reachability in the AS-level graph, *i.e.*, many physical paths are not visible by the inter-domain routing protocol for connectivity recovery during Internet outages. We call the invisible connectivity at the routing layer, which can be quickly restored for recovering routing failures by simple configurations, as the potential routing diversities. In this paper, we evaluate two kinds of potential routing diversities, which are recognized as Internet eXchange Points (IXPs) participant reconnection and peering policy relaxation. Using the most complete dataset containing AS-level map and IXP participants that we can achieve, we successfully evaluate the ability of potential routing diversity for routing recovery during different kinds of Internet emergencies. Encouragingly, our experimental results show that 40% to 80% of the interrupted network pairs can be recovered on average beyond policy-compliant paths, with rich path diversities and a little traffic shifts. Thus, this paper implies that the potential routing diversities are promising venues to address Internet failures.**

***Index Terms*—Internet reliability, failure recovery, routing.**

I. Introduction

AS the Internet becomes a critical infrastructural component of our global information-based society, any interruption to its availability can have significant societal impacts. Thus, when listing its requirements on the Internet, the GENI initiative [1] states that: “any future Internet should attain the highest possible level of availability”. Unfortunately, despite of the remarkable availability and responsiveness demonstrated by the Internet in most cases, they still need a substantial improvement when a disaster/emergency[1] strikes [2], such as 911 terrorists attack [3], Taiwan earth quake [4] and fiber cut in San Francisco area [5]. Take the Taiwan earthquake that struck on Dec. 26, 2006 as an example. This earthquake has caused significant damages to the Internet and even one week after the Taiwan earthquake, the number of outage Internet networks was still in the order of thousands [6].

The Internet is a network of Autonomous Systems (ASes) and the Border Gateway Protocol (BGP) is the de facto inter-domain routing protocol among ASes. The understanding of BGP’s dynamic and Internet topology is crucial for the analysis on Internet availability and resilience. In [7], the authors highlighted the bad impact of BGP’s dynamics on the data plane where data packets suffer loss, delay, and reordering. In [8], the authors performed on k-shell analysis and the resiliency of the Internet to top shell (core) disconnections and discussed how paths traverse the Internet shells and quantify the damage of core failures. In [9], the authors investigated the impact of factors such as policies, topology, IGP weights on routing stretch and diversity. Recently, there are several literatures investigating the Internet resilience enhancement on the BGP routing layer [10]–[13]. These techniques only focus on exploring and utilizing the policy-compliant path that satisfies the inter-domain policy of BGP. Therefore, the potential of all these work is inherently constrained by the underlying resilience of BGP-based Internet routing structure which is analyzed in [14]. In [14], the authors made a detailed measurement analysis about the Internet’s resilience to failures based on BGP and their results revealed that today’s Internet is vulnerable. For example, they disclosed that 32% of all the ASes are vulnerable to the failure like a single critical customer-provider link cut, and up to 93.7% of the reachability of Tier-1 ISP’s single-homed customers has been lost due to the failure like Tier-1 network depeering. Please note that these results are also applied to all the existing work [10]–[12] that utilizes the policy-compliant paths for failure recovery as the upper-bound of capabilities.

Given the insufficiency of the existing work, we continue to seek other solutions to help enhance the Internet failure recovery. In order to excavate more available resources for help, we need to profile the requirements for potential assistance. Consider a router A belongs to the victim AS, and another router B belongs to a survival AS in the failure. We say that

Manuscript received September 28, 2011; revised February 7, 2012; accepted Match 5, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was R. Stadler.

This work was in part presented at INFOCOM 2010 (Mini-conference) and NANOG 49.

C. Hu is with the MoE Key Laboratory of Intelligent Networks and Network Security, Department of Computer Science and Techonolgy, Xi’an Jiaotong University, Xi’an, Shaanxi, China (e-mail: huc@ieee.org).

K. Chen is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology (e-mail: ckg.com@gmail.com).

Y. Chen are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA (e-mail: ychen@northwestern.edu).

B. Liu is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China (e-mail: liub@tsinghua.edu.cn).

A. V. Vasilakos is with the Department of Computer and Telecommunications Engineering, University of Western Macedonia, GR 50100 Kozani, Greece (e-mail: vasilako@ath.forthnet.gr).

Digital Object Identifier 10.1109/TNSM.2012.12.110191

[1] When referring to “disaster” or “emergency”, we mean an incident which causes a large number of disconnections among ASes for a long period.

there exists a candidate emergency recovering connectivity between A and B if the following two basic features are satisfied.

- **Fast on-demand connection.** There already exists a physical link connecting A and B.
- **Reachability.** Router B can reach at least one IP prefix belonging to the set of recovering destination IP prefixes defined by the victim AS.

The BGP-based Internet routing protocol is a policy-driven paradigm under which the physical connectivity does not necessarily translate into network reachability, causing many physical paths are not visible to the victim network in a failure emergency. To further make use of these physical paths, in this paper, we exploit and evaluate the *potential routing diversities*, which are one step beyond the scope of BGP policies. Specifically, at least two kinds of inter-domain links can be used for emergency recovery via simple configurations. They are:

- **Setting up BGP sessions between IXP participants.** Internet eXchange Point (IXP) is a physical infrastructure that allows multiple ASes to exchange Internet traffic between their networks. Since the co-located ASes in a same IXP have routers geographically nearby and physically interconnected by switches [15], a new BGP session between two ASes (which are physically connected through switch but not connected through BGP previously) in the IXP can be easily set up.
- **Policy relaxation between neighboring ASes.** By relaxing the policy restrictions between victim AS and its neighboring survival ASes in an emergency, the previously policy-prohibited routing can be reused and the outage could be potentially recovered.

While the idea of policy relaxation has been mentioned in [14], setting up new BGP sessions between IXP participants to rescue the failure emergencies is a novel idea proposed in this paper which is *the first contribution*.

To discover and utilize these potential routing diversities is an ambitious endeavor. As a first step towards this goal, we aim to answer in this paper how much connection outage can be potentially recovered using the possible invisible physical paths in the inter-domain routing layer in addition to the BGP self-recovery. To the best of our knowledge, this question has not been answered before. We have illustrated several preliminary results in our previous work [16], and this paper presents more comprehensive experiments and evaluation results. This is *the second contribution* and also the major contribution of this paper.

We organize the paper as the following. In Section II, we specify the two kinds of potential routing diversities in details. In Section III, we manage to leverage an AS graph to date containing 31,845 nodes (ASes) and 142,970 links for the analysis on the effectiveness of potential routing diversities. The links are extracted from a large-scale 541M traceroute measurements from 580K end hosts, as well as all publicly available BGP tables. We also describe in this section the IXP information collected from several available datasets. In Section IV, we check the severe failure models where critical links or a large number of links are cut[2]. The evaluation results shed new light on the inter-domain routing resilience using potential routing diversities. In the remaining of the sections, we discuss the limitations of this paper in Section V, review the related work in Section VI and summarize the paper in Section VII.

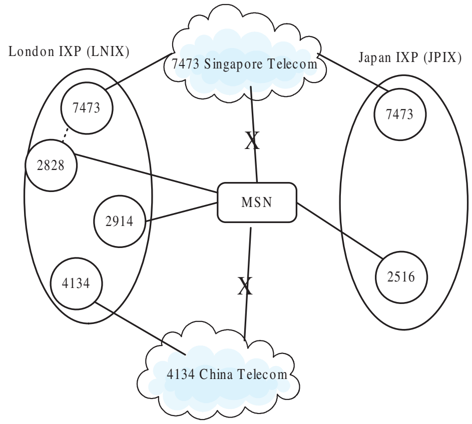

Fig. 1. Reconnecting routers in IXPs to achieve network reachability.

II. Exploring Potential Routing Diversity

In this section, we describe the two potential routing diversities that could be utilized for network reachability recovery in case of Internet emergencies.

A. Reconnect IXP participants

An IXP is a physical infrastructure that allows different ISPs to exchange Internet traffic between their networks by means of mutual peering agreements. In an IXP, although the co-located ASes have their routers geographically nearby or have even physically interconnected by Layer-2 switches (or layer-2 cloud, layer-2 switching fabric) [15], a BGP session is not necessarily and naturally established between any two ASes unless the corresponding ASes have business relationship between them.

In case of Internet failure emergencies, some of connectivities to the lost networks can be recovered by establishing new BGP sessions between (selective) co-located ASes. A new BGP session between two ASes (which are physically connected through switch but not connected through BGP previously) in an IXP can be easily established. In order to setup a BGP connection between the router of victim AS and the router of survival AS, each of them only needs to have a “neighbor *ip-address* remote-as *AS-number*” entry in its configurations. We envision that the IP addresses and the AS numbers of the participants in an IXP can be obtained through the third party who manages it. Even an automated mechanism over the layer-2 switching fabric is a feasible alternative in case the human communications would delay on recovery process. These design details are not the focus of this paper and are left as our future work.

[2] Please note that the BGP is resilient to small-scale failure, *i.e.*, teardown of a single link in lower tier, as demonstrated in previous work. So we do not repeat these experiments

TABLE I
EXPORT POLICIES OF INTER-DOMAIN ROUTING.

	peers	providers	customers	siblings
To peer			✓	✓
To provider			✓	✓
To customer	✓	✓	✓	✓
To sibling	✓	✓	✓	✓

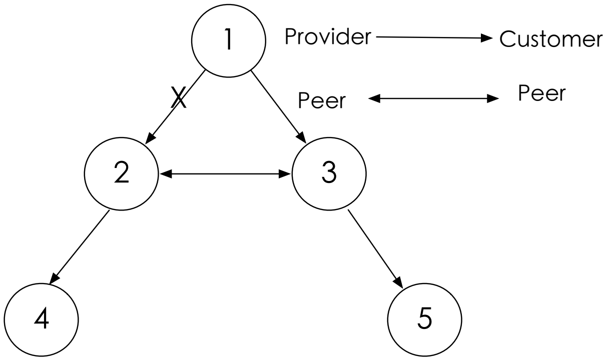

Fig. 2. Relaxing routing policy to achieve network reachability.

To describe how IXP participants reconnection can be used as the potential routing diversity, we use an example scenario in Japan IXP (JPIX) and London IXP (LNIX) during the Taiwan earthquake incident. As shown in Fig. 1, some ASes in these two IXPs is selected for illustration. In the figure, the big ovals stand for IXPs and small cycles are ASes in the IXPs. During the disaster, Singapore Telecom (AS7473) and China Telecom (AS4143) lost their connections to MSN, but AS2828, AS2914 in LNIX and AS2516 in JPIX are still able to access MSN. To recover the traffic from AS4134 to MSN, we could create a new provider-customer link between AS4134 and AS2914 (or AS2828) in LNIX and then the traffic would traverse AS 2914 (or AS 2828) to reach MSN. To recover the traffic from AS 7473 to MSN, we could set up a new provider-customer link between AS7473 and AS2914 in LNIX (or between AS7473 and AS2516 in JPIX).

B. Peer relaxation

The exporting policy [17] of BGP routing obeys the following two rules as shown in Table I: 1) each AS exports to its providers/peers its own routes and those learned from its customers or siblings; 2) each AS exports to its customers/siblings its own routes and any routes learned from others. According to the exporting policy, the valid AS paths follow the "valley-free" rule [18] and a peering link or customer-to-provider link never carries traffic from another peer or provider. As a result, the routing diversity of BGP has been restricted and physically connected paths may not be visible to the victim network in a failure emergency. For the example in Fig. 2, if link 1-2 is broken then AS2 and AS4 are disconnected from AS1 although the underlying physical path 4-2-3-1 exists. This is because link 3-1 is a customer-to-provider link, it will not be used to carry traffic from a peer, *i.e.*, AS2.

If the exporting policy on some links could be relaxed, the previously policy restricted AS path(s) can be utilized to carry affected traffic when Internet emergency happens and the physical connections beyond the BGP policy are fully exploited in this way. Again, we use the example in Fig. 2 to present the idea of policy relaxation. If the relationship on link 2-3 is changed from peering to customer-to-provider, link 3-1 can now carry traffic from AS2. Through this way, the policy-relaxed AS path 4-2-3-1 works to carry traffic from AS4 and AS2 to AS1, which is not visible from the original BGP policy routing.

In principle, both peering links and provider-to-customer links can be relaxed into customer-to-provider links, however, in this paper, we only focus on evaluating the potential of relaxing peering links because peer-peer links usually have more bandwidth than the links to its customers and relaxing a customer-to-provider link to a provider-to-customer is less likely to happen in reality.

III. DATASET, METHODOLOGY AND MODEL

A. AS topology

A complete Internet AS topology graph is definitely useful to our evaluations on potential routing diversity. To this end, we use both BGP data and traceroute data to build an AS graph that contains 31845 AS nodes and 142970 links.

BGP data: First, we use AS links from UCLA IRL lab [19] as a base, which are collected from route servers, looking glasses, and Internet Routing Register [20]. Since this data set does not provide BGP AS paths and information from newly added vantage points, we also collect the BGP data from 790 BGP speaking routers in 438 unique ASes. Specifically, we combine several BGP feeds: Routeviews [21] collected at route-views.oregon-ix.net, which is the most widely used BGP archive so far, 6 other Oregon route severs and 16 route collectors of RIPE/RIS [22].

Oliveira et al. [15], [23] pointed out that ten months of the public view data should be enough to cover all the customer-provider links and upstream peering links in the Internet AS graph. According to this, we use ten months of data gathered between Dec 1, 2007 and Sep 30, 2008. However, the graph drawn from only 10-month BGP data still misses a lot of low-tier peering links as mentioned in [15]. Therefore, we further use traceroute data to find hidden peering links.

Traceroute data: The traceroute data are collected by P2P users located in 580,000 host across the 5,500 ASes at the same time as we collect BGP data, and to the best of our knowledge, this is the largest scale traceroute measurement when we started the work in this paper, which consists of 541,023,742 measurements over 6.2 billion hops [24]. Using a set of heuristics, we have extracted about 23,000 links that are complementary to the BGP AS links. The heuristics and traceroute AS links are validated with real ISPs. Please refer to [24] for the details of the methodology and heuristics.

Network classification: The Internet structure is considered to be loosely hierarchical. There are methodologies to classify the Internet tiers according to the degree of each individual AS, or the number of prefixes originated by the AS, or the distinct AS paths seen from the AS, etc. However, without accounting for the AS contractual relationship, these heuristics may be misleading. For example, the degree for an AS may include a mixed set of neighbors including providers, peers,

TABLE II
STATISTICS OF ASES ON EACH TIER.

Tier-1	large AS	media AS	small AS	stub AS
9	281	1644	4642	25269

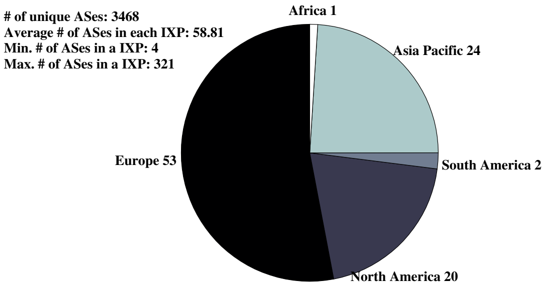

Fig. 3. IXP statistics.

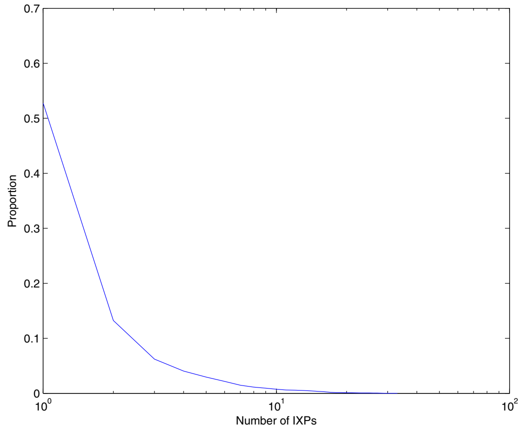

Fig. 4. The proportion of non-stub ASes that has presented in at least x IXPs.

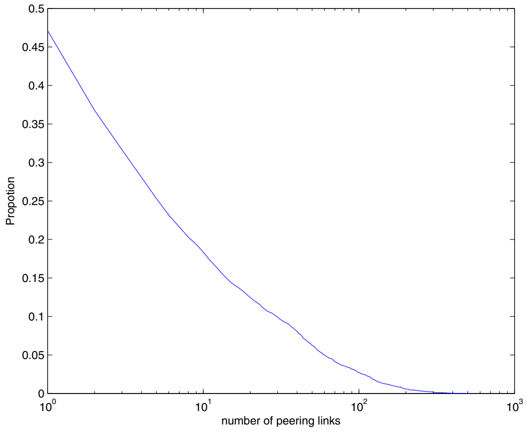

Fig. 5. The proportion of non-stub ASes that has at least x peering links.

or customers. Algorithms based on the number of prefixes and distinct AS paths may be too coarse since prefixes vary significantly and route aggregations happen everywhere. To mitigate these limitations, we apply the state-of-the-art method proposed in [15] which uses the number of downstream customer ASes to classify AS hierarchy. Besides 9 well-known Tier-1 ASes, "large AS" is the AS that has more than 50 customers, "media AS" is the ones with 5-50 customers, "small AS" has less than 5 customers, and "stub AS" is the AS which does not provide transit service to any other AS. The number of ASes in each category is shown in Table II.

B. IXP dataset

There are several sources, such as Packet Clearing House [25], Peeringdb [26], and Euro-IX [27], each of which maintains a list of IXPs, as well as their participants. While there are more than 200 unique IXPs worldwide in these sources, we intentionally pick 100 of them for our study according to their scale and geographical locations. The number of participant ASes in other IXPs is quite small and the contributions of them are minor. So we exclude them from the experiments in this paper.

The statistics of IXPs is listed in Fig. 3. There are 3468 unique ASes (all of these ASes are not stub ASes) presented in our IXP data set, *i.e.*, 52.7% non-stub ASes are involved in at least one IXP. One AS can be involved in a number of different IXPs. More IXPs an AS participants, larger probability it could be recovered by the IXP helpers. In average, one IXP member can be in 1.70 IXPs and in the best case, one AS appears in 33 IXPs. As shown in Figure 4, 52.7% of the unique IXP members are only in one IXP and only 6.3% of the non-stub ASes can be in more than three IXPs.

C. Policy inference

We infer the business relationships between ASes based on the Partialness To Entireness (PTE) algorithm proposed by Xia [28] and most AS links are classified as one of three kinds of relationships: *customer-provider* links, *peering* links, and *sibling* links. By improving the seminal work by Gao [18], the PTE method is considered to outperform most other approaches [29]. The basic idea of PTE algorithm is to infer the entire AS relationships from partial information. The algorithm consists of two major components. One is to filter non-valley-free paths (use pre-fetched partial AS relationships to examine the valley-free property for each AS path, and then remove the AS paths which cannot be valley free from the data), and the other is to infer AS relationships from partial information (define three inference rules on an AS path and one refreshing rule based on valley-free property of AS paths).

Since peering links are used for peer relaxation, we analyze the distribution of peering links here. Without stub AS, the average number of peering links for an AS is 5.35 and an AS could peer with 503 ASes at most in our dataset. Fig. 5 demonstrates the proportion of non-stub ASes that has at least x peering links. Among all the non-stub ASes, 47.1% have at least one peering link and 18.4% have more than 10 peering links.

TABLE III
STATISTICS OF THE AS TOPOLOGY IN OUR STUDY.

	# nodes	# links	# customer-provider	# peering	# sibling
with stub AS	31845	142970	94500 (66.1%)	43709 (30.6%)	4761 (3.3%)
without stub AS	6576	78090	38580(49.4%)	35151(45.0%)	4359(5.6%)

TABLE IV
LIST OF TIER-1 ASES.

ASN	174	209	701	1239	2914	3356	3549	3561	7018
AS	PSINET	Qwest	UUNET	SPRINT	Verio	Level 3	Global Crossing	Cable & Wireless	AT&T

D. Failure models

We measure the proposed routing diversities to understand their potential on recovering the Internet failures. We classify the failures according to the types of the failed links, *i.e.*, teardown of peering links, provider-customer links (access links) and mixed link types.

- **Tier-1 depeering.** A depeering[3] could be caused by contractual reasons, mis-configurations or physical damages. As evidenced by contractual disputes between Cogent and Level 3 [14], [30], Internet connectivity can be significantly affected by the depeering over a peering link between two tier-1 ASes.
- **Access links teardown.** Provider-customer links connect the networks in different tiers of the Internet, which contribute to the major reachability. Tier-1 ASes construct the core of today's Internet and are the top service providers.
- **Regional failures.** Regional failure breaks several peer links and provider-customer links, which can be caused by a regional emergency.

IV. EVALUATION RESULTS

A. Evaluation metrics

When failures happen, they disrupt the traffic that traverses the failed network components and break the connectivities between a number of $< src, dst >$ pair, which is called *lost* $< src, dst >$ *pair*. When BGP converges, the connectivity of some lost $< src, dst >$ pairs can be recovered; however, there are still many lost$< src, dst >$ pairs whose connectivity are still not restored after BGP convergence and such pairs are named as *non-reachable* $< src, dst >$ *pairs*. We introduce the potential routing diversities for recovery of the non-reachable $< src, dst >$ pairs and evaluate its effects using the following methods.

- **IXP reconnection (IXP).** We first check whether a victim is in any of the IXPs in our IXP dataset as mentioned before. If it is in at least one IXP, the reachability of the ASes in all the co-located IXPs to the recovering destination is checked. Again, the reachability could be recovered if at least one of the participants located in the same IXP is able to access the destination.
- **Peer Relaxation (PR).** We validate this kind of potential routing diversity by investigating the reachability of all the peers of the victim. As long as at least one of its peers reaches the recovering destination, the connectivity of the victim to the destination could be recovered by relaxing the peer relationship.

Furthermore, we define the following metrics to quantify the potential:

- **Recovery ratio** is the ratio between the number of non-reachable $< src, dst >$ AS pairs that can be recovered by using potential diversities and the total number of the non-reachable $< src, dst >$ AS pairs.
- **Path diversity** of a $< src, dst >$ pair stands for the number of parallel paths that do not share links and it can be formally defined as the minimal number of links (including the link itself) that must be removed in order to disconnect the two endpoints of the pair. This is used to show the path redundancies between the source AS and the destination AS.
- **Shifted path.** The traffic from the failed link will shift to a new link after recovery. Since it is impossible to get the traffic metrics over each two ASes, we estimate the amount of traffic over a certain link as the number of the shortest valley-free paths that traverse the link as [14] did. The number of the shifted path is evaluated in the experiments.

B. Depeering

In this section, we check the potential routing diversities during the tier-1 depeering since it is the worst case for depeering failures as indicated in [14]. Table IV illustrates the nine well known tier-1 ASes, which are used to generate tier-1 depeering scenarios for experiments. Totally 36 experiments are performed, and in each experiment, we: 1) assume one peering link between two tier-1 ASes to be down and then find the lost $< src, dst >$ pairs which are originally connected through this link; 2) seek alternative valley-free routes for the these pairs and the $< src, dst >$ pairs that are still not reachable are picked for our evaluation of potential resources; 3) check the metrics described above via peer relaxation, IXP participant reconnection and both.

1) Recovery ratio: Among the 36 experiments, with only peer relaxation, the minimum, mean and maximum recovery ratio are 0.44, 0.65 and 0.86; with only IXP participant reconnection, the minimum, mean and maximum recovery ratio are 0.28, 0.48 and 0.64; and with both peer relaxation and IXP participant reconnection, the minimum, mean and maximum recovery ratio are 0.63, 0.78 and 0.92. The absolute number of non-reachable pairs is 31456 (with a maximum of 144516 and a minimum of 1996). The cumulative distribution

[3] Depeering: if one or both networks in a peering relationship believes that there is no longer a mutual benefit, they may decide to cease the free exchange of traffic.

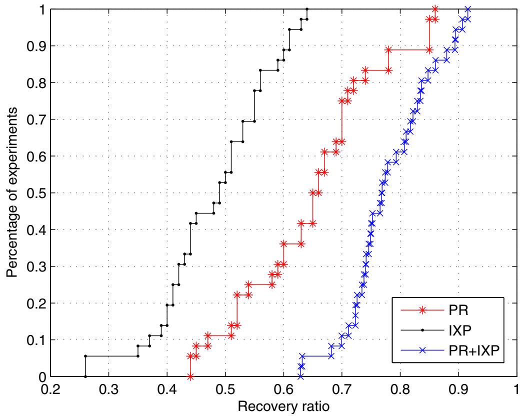

Fig. 6. CDF of the *recovery ratios via PR, IXP,* and *PR+IXP* for tier-1 depeering.

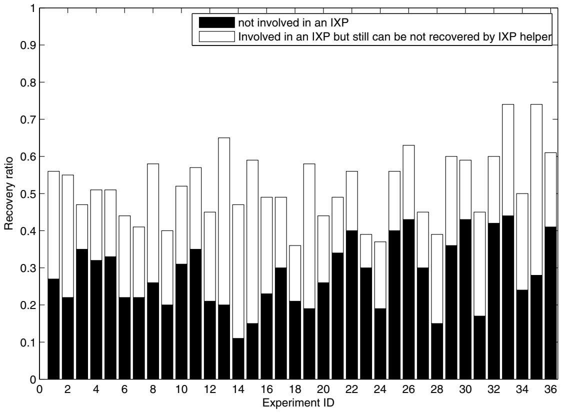

Fig. 7. Details of not recovered AS through peer relaxation for tie-1 depeering.

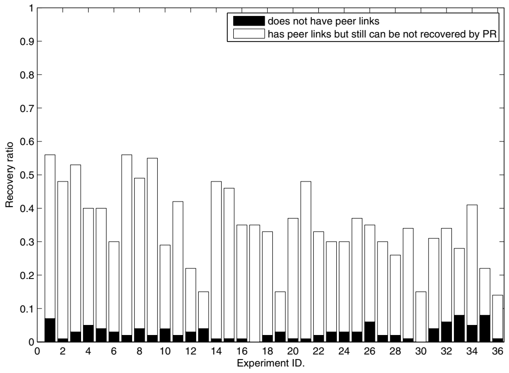

Fig. 8. Details of not recovered ASes through IXP participant reconnection.

of the percentage of experiments to recovery ratios via PR, IXP, and PR+IXP of the above experiment are depicted in Figure 6. The figure shows the proportion of experiments where recovery ratio is less than a specific value. We can see that a significant portion of the AS pairs can be recovered with PR or IXP. The recovery ratio via PR is larger than those via IXP because the number of peering links that can be used for peer relaxation is much larger than the number of IXP participant reconnection.

Next, when a victim AS is involved in an IXP or has peering links, we check how likely its connectivities can be recovered. In Fig. 7 and Fig. 8, we look into the non-reachable pairs that are failed to be recovered by peer relaxation and IXP participant reconnection, and inspect how many of them still cannot be recovered. In both figures, the black part presents the proportion of the victim ASes that do not have peering links or do not appear in any IXP; while the white part is the proportion that no alternative routes are found by peer policy relaxation or IXP participant reconnection. We observe that, 1) among all non-reachable pairs that cannot be recovered by peer relaxation, only 9.4% of them do not have peering links (*i.e.*, 90.6% of them do have peering links but these peer links could not help during failures); 2) among all non-reachable pairs that cannot be recovered by IXP participant reconnection, 54.3% of them do not appear in any IXP (*i.e.*, 45.7% of them is in IXPs but could not find IXP participant reconnection). The number indicates that the probability to get potential connectivity if a victim AS is involved in an IXP is much larger than the case that it has peer links.

Note that if two peered ASes are both single-homed to a same tier-1 AS, in case of tier-1 peering link, they cannot help each other. This explains why a lot of victim ASes cannot find potential routes from its peers as discussed before.

2) Path diversity: 36 experiments (each of which breaks two tier-1 ASes) are performed to study the improvement on the path diversity metric of the recovered $< src, dst >$ pairs. The path diversities are all 0 for the non-reachable pairs before the recovery. After introducing the potential connectivities, the path diversity of each recovered $< src, dst >$ pairs increases. Take one simulated experiment as example where AS1239 and AS2914 depeer. The average path diversity of the recovered pairs by PR and IXP is 6.2, which means there are 6.2 parallel paths (on average) can be chosen by a victim AS to reconnect to the destination. Among all the 36 experiments, the average path diversity is 3.6, the minimum path diversity is 1.9 and the maximum path diversity is 6.3. The distribution of the average path diversity in each experiment is depicted in Fig. 9. In most of the experiments, the path diversity is 2-3 (12 experiments, 33.3%) or 3-4 (14 experiments, 38.9%). The existence of such multiple parallel paths gives the chance to avoid possible congestion caused by traffic moving from links to links.

3) Shifted path: The average increased number of paths traversing a link, *i.e.*, the average number of shifted path, varies from 3.75 to 17.2 in all the 36 possible scenarios of tier-1 AS depeering disaster. We also observe that the number of the shifted paths to some links could be very large and the maximum number of the shifted paths to a link varies from 174 to 4217 in the 36 experiments. Fortunately, only a very small number of paths are shifted over most of the links. Fig. 10 shows the percentage of links with the number of shifted paths, when AS1239 and AS2914 depeer. In this case, the number of shifted path over 86.3% of the links is no more than 4.

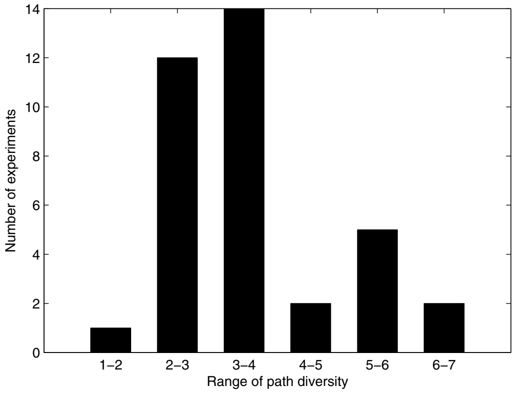

Fig. 9. Distribution of path diversity among 36 experiments and each experiment mimics the case that breaks the link between two tier-1 ASes.

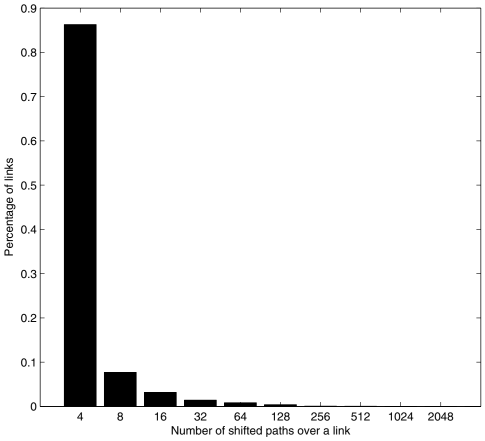

Fig. 10. Histogram of the number of shifted paths if AS 1239 and AS 2914 depeer.

C. Access link teardown

Today's core Internet consists of a small number of tier-1 ASes acting as the top transit providers. Damages of several provider-customer access links belonging to its tier-1 AS could cause the disconnections of a large number of customers and "grand-customers" from Internet. Therefore, we mimic the failure emergency of multiple provider-customer links by randomly picking access links between tier-1 (Table IV) to set its customers to be down simultaneously.

Similar as the steps in tier-1 depeering experiments, existing routing diversity by BGP valley free feature is first used to recover the lost AS pairs caused by the disconnection of multiple provider-customer links. And then the potential connectivity diversity is evaluated for the $< src, dst >$ pairs that can not be recovered by BGP routing.

1) Recovery ratio: The number of non-reachable $< src, dst >$ pairs in this failure type is relatively smaller than the ones in tier-1 depeering example. Even there are 30 failed access links to each tier-1 AS, the number of non-reachable $< src, dst >$ pairs is 1835 on average (with a maximum of 6515 and with a minimum of 240). The recovery ratios via PR, IXP and PR+IXP are illustrated in Table V, and we can recover 1/3 to half of the non-reachable $< src, dst >$ pairs, even in the face of severe failures, *e.g.*, 30 customer-provider links are down simultaneously. While PR and IXP provide similar potential routing diversities in this failure type, IXPs are more efficient than peering links for failure recovery, since the percentage of victim ASes that are not in any IXP is larger than the victims that do not have peers.

2) Path Diversity: The path diversities of the recovered ASes pairs are also checked under this failure scenario. The average path diversity is increased by 4.64 with peer relaxation and IXP participant reconnection when 10 provider-customer links are down. It does not show an obvious decrease in the increased path diversity when the number of broken links increases. The path diversity is 4.54 on average for the case that 20 provider-customer links fail.

3) Shifted path: The average number of shifted path when 10, 20 and 30 links are damaged are 3.4, 4.0 and 4.2, respectively. It is a slight increasing trend with the increase of the number of the affected links. Although the number of shifted paths over some links could be as large as thousands, the number of shifted paths is less than 4 over about 87.9% to 98.3% of the links in different experiments with different number of simulated broken access links.

D. Regional Failure

In this section, we check the potential resilience when regional failure occurs. Specifically, we investigate the potential connectivity improvement to Taiwan earthquake incident stricken in [6].

1) Recovery ratio: There is no pubic data indicating the exactly destroyed links during the Taiwan earthquake. To simulate the disaster and study the above nine large victims, we randomly remove 50% of the links for the nine heavily affected ASes in the disaster as indicated by [6]. These ASes are AS4134, AS4755, AS4761, AS4795, AS4837, AS9498, AS7473, AS9929 and AS24077. By checking BGP AS paths, we collect the AS pairs that traverse these removed links as the lost $< src, dst >$ pairs in the experiment. Then, we check whether there is an alternative path (*i.e.*, valley free) to recover the lost $< src, dst >$ pairs. The still non-reachable pairs are used to evaluate the potential by introducing peers and IXP participant reconnection. In the simulation, we observe that the number of non-reachable pairs caused by the failure of different victim varies from 644 to 103236.

The recovery ratios via PR, IXP and PR+IXP are shown in Fig. 11. Each group of the three bars indicates the recovery ratio to the disconnections caused by the deleted links belongs to one of the victim AS. The average recovery ratios via PR and via IXP are 52% and 56%, respectively. Utilizing both of them, the average recovery ratio is increased to 70%. Among the not recovered non-reachable $< src, dst >$ pairs through IXP participant reconnection, 63% of them are not involved in any IXP; and among the not recovered non-reachable pairs through peer relaxation, 25% of them do not have peer links.

2) Path diversity: According to the feedback from a large ISP in China, the accesses to MSN, Google and DNS roots were strongly required by its customers during the Taiwan earthquake incident. To evaluate the resilience improvement

TABLE V
POTENTIAL ROUTING DIVERSITY BY PR OR IXP WHEN SEVERAL PROVIDER-CUSTOMER LINKS ARE DAMAGED.

Links down	recovery ratio via PR	No peer	Recovery ratio via IXP	Not in IXP	Recovery ratio via PR+IXP
10	35.8%	10.7%	32.7%	26.2%	43.7%
20	27.8%	10.9%	27.6%	29.33%	38.1%
30	26.8%	9.4%	26.0%	29.8%	37.0%

TABLE VI
PATH DIVERSITY IMPROVEMENT. "PD" IS SHORT FOR PATH DIVERSITY AND "IR" STANDS FOR IMPROVEMENT RATIO BETWEEN INCREASED PATH DIVERSITY AND THE ORIGINAL DIVERSITY.

Des.	BGP only	BGP+PR		BGP+IXP		BGP+PR+IXP	
	PD	PD	IR	PD	IR	PD	IR
MSN	24.4	27.6	13.0%	32.1	31.6%	34.1	39.8%
Google	14.9	17.6	17.9%	21.1	41.8%	23.3	56.5%
DNS	105	112.8	7.4%	123.6	17.7%	127.2	21.1%

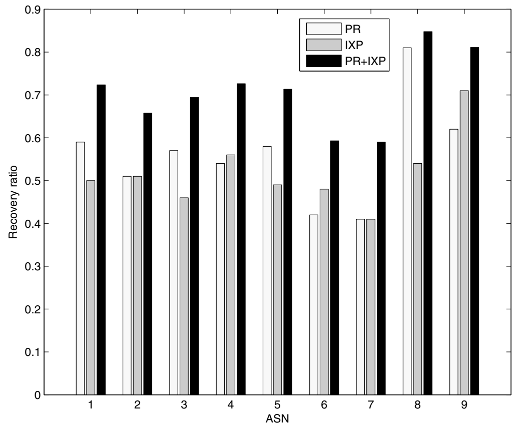

Fig. 11. The recovery ratio by potential connectivity diversity beyond BGP for Taiwan earthquake incident.

via potential routing diveristies, we try to capture the changes on the number of the unique paths (*i.e.*, the path diversity) to these destinations from the 9 heavily affected ASes with and without the utilization of potential resources. Table VI illustrates the results of path diversity improvement based on the potential connectivities. All the valley free paths without sharing any shared risk link are first collected for each $< src, dst >$ pair. The average number of unique paths of the 9 sources to each of the destination are listed in the column "BGP". Second, all the peering links directly connected to the source are relaxed to be provider-customer links. The path diversity is checked again and the results are in the column "BGP+PR". Third, by adding new BGP sessions between each source AS and its potential IXP participant reconnection, the diversity is shown in "BGP+IXP" column. The results from both peer relaxation and IXP participant reconnection are illustrated in the last column as "BGP+PR+IXP". In the table, we can see that,

- The link connectivity to DNS is much larger than the other two. This is because there are 13 DNS servers, and the link connectivity to DNS is the summation of the one to each of the DNS server;
- Solely leveraging IXP participant reconnection can achieve more improvement than only relaxing the peering links;
- Since MSN, Google and DNS servers are hot spots of Internet, the connectivity is already quite large. However, by peer relaxation and IXP reconnections, it still provides non-trivial increases in path diversity by 21.1% to 56.5%.

In addition, we study the increased path diversity to all the destinations. The increased path diversity for the non-reachable pairs, sourced from the nine ASes aforementioned is illustrated in Table VII. The value listed in each column is the mean over all the recovered non-reachable pairs. It is shown that at least two parallel paths could be found in average through peer relaxation and IXP participant reconnection.

3) Shifted paths: The number of the shifted paths is shown in Table VIII. The first column indicates the failures caused by the links to a certain AS. The second column is the average number of shifted paths, while the third column lists the maximum number of shifted paths. To view the distribution of the shifted paths, the fourth column demonstrates the percentage of links, over which the number of shift paths is less than 4. Obviously from the table, we observe that, 1) the average number of the shifted paths is small; 2) there are heavily used links which absorbs a large number of shifted paths; 3) however, the number of such links are quite small and less than 4 shifted paths are over 77.7% to 90.1% links.

E. Summary of Main Findings

Through the above experiments, we summarize the following main findings.

- **Potential routing diversities significantly improve the resilience.** Even in heavy failure models, approximately 40% to 80% of all non-reachable source-destination ($< src, dst >$) pairs that cannot be recovered after BGP's convergence are now reconnected by using the potential routing diversities.
- **Peer relaxation is easier, and IXP participants is more capable.** 47.1% of the non-stub ASes have peering links in the AS graph we studied, however, most of the peer links can not used to recover the failures in some severe cases. Compared with peering relaxation, only a few ASes are involved in at least one IXP, but the chance for a victim AS to recover its connectivity through IXP participant reconnections is relatively larger.
- **Potential routing diversities themselves have redundancy.** Several parallel paths that do not share any link,

TABLE VII
INCREASED PATH DIVERSITY FOR TAIWAN EARTHQUAKE FAILURE.

ASN	Increased path diversity
4134	3.4
4755	2.2
4761	2.3
4795	2.5
4837	2.5
9498	2.4
7473	2.4
9929	2.8
24077	4.2

TABLE VIII
SHIFTED PATH FOR TAIWAN EARTHQUAKE FAILURE.

ASN	average	max.	less than 4
4134	6.2	413	89.6%
4755	6.0	790	82.4%
4761	7.6	529	84.4%
4795	6.9	261	78.7%
4837	7.2	259	77.7%
9498	7.9	652	79.3%
7473	9.1	3682	81.8%
9929	3.8	55	83,8%
24077	3.9	43	90.1%

which can recover inter-domain routing through potential routing diversities, is discovered even in severe failures.

- **Impact on traffic path shift can be controlled.** No more than 25% of the links absorb/disperse more than 4 traffic flows in the $< src, dst >$ granularity. Thanks to the redundant potential routing diversities mentioned in the above item, a smarter policy on path selection could further reduce the path shift.

V. LIMITATIONS AND DISCUSSIONS

Considering that the existing Internet resilience is not sufficient for emergency recovery, in this paper we demonstrate the feasibility to exploit the *potential routing diversities* during the Internet failures. Our intention to extend the routing diversities is not meant to replace existing approaches for recovering the Internet failures; rather, it is complementary to existing work in that it focuses on evaluating the potential resources for advanced recovery mechanisms.

While promising, there are limitations within our work. The quality of the Internet AS topology is very important to evaluate the potential routing diversities addressed in this paper. Although we have used the most complete AS topology graph to date, we still do not know how many links are missing from our graph. This potentially influences our measurement results. Furthermore, the accuracy of inferred AS relationships has impact on the peering relaxation and validity of valley-free inferred AS paths. We leveraged the PTE algorithm proposed by Xia [28] to infer the AS relationship. While it is considered to outperform most other approaches [29], the results may not be 100% correct. Such limitations on topology and relationship inference are shared by all the existing work since available control/routing- and data-plane measurements both suffer from some kind of biases [31]. Also, a failure that disconnects two networks in our failure model, may actually disconnect multiple IP links at the same time if they share the same physical infrastructure. However, we cannot know the physical dependencies between different links. That is other limitation of this paper.

We have evaluated the ability of the *potential routing diversities* in this paper, and the idea of exploiting these routing diversities is deployable. In [32], we present the design of IER (Internet Emergency Response) system to substantially realize speedup the recovery process of the Internet availability after an Internet emergency, based on the routing diversity. We introduce a detailed IER framework, which contains three main modules: candidate helper AS identification, resource allocation and network reconfiguration. In the first module, we build a communication channel among the routers. Over the channel, the affected ASes broadcast their desired resources (e.g., lost destinations) and the helper ASes advertise what they can help. This is the first thing that to find potentially help to reach the lost destinations, when an emergency happens. Once the help information is advertised to the affected ASes, we enter the resource allocation module. The task here is to accommodate the demand-and- supply between affected ASes and helper ASes. Note that the resource allocation involves practical considerations. For example, resources are not free and helper ASes may charge money for their help. From an affected ASs perspective, it may want to reach as many lost destinations as possible with fewer new AS contracts. From an helper ASs perspective, it may want to sell as many resources as possible with fewer new contracts. After the resources being arranged, we need to reconfigure the routers accordingly. However, improper reconfiguration would cause the newly established contracts to carry unexpected traffic. To this end, we identify and analyze the root causes for the unexpected traffic, and then propose the solutions to mitigate the problem in the IER system. Furthermore, we evaluate the effectiveness of the IER system in [32] using synthetic data generated from the realistic Internet AS topology and IXP dataset, and the results have demonstrated that the process is fast and is able to deliver reasonably good recovery rates in a series of settings, for example, in a major emergency, it can figure out how to recover about 2.4 million disconnected AS pairs within 11 seconds.

ISP marketing is complex, but the network owners have clear incentives to manage failures well, because unmanaged failures can cause severe service disruptions and lead to significant financial and reputation damages [33]. Building dedicated redundant networks or setting up backup arrangements ahead of the failure requires significant investment. It is impossible to predict when and where a failure will occur, the coverage and gain of such investment on redundancy has essential limitations. Also, the serious emergency is not frequent and the risk for different ASes is comparable in a long run. Using potential routing diversity, an affected AS only "borrows" connectivity from others on demand during the failure, therefore, this framework pools the resources of multiple ASes together for mutual backup to improve network reliability at low cost. Similar cases are also activated in other area of our life, for example, airline industry uses other airlines ights to transport their customers when their own aircrafts are not available to fly. In addition, the help process is relatively short and affects the helper ASes temporarily, and the helper

ASes can also make profit by helping the affected ASes [32].

VI. Related Work

Although the routing on today's Internet has demonstrated remarkable availability and responsiveness on average, there is a lack of resilience to failure emergencies. There are a spectrum of literatures studying the Internet reliability or resilience under failures. While they all have made good efforts on advancing the Internet reliability, each of them also has its own limitations.

Solving Intra-AS failures. In paper [33], the authors propose a solution framework called reliability as an interdomain service (REIN) to improve the redundancy of a single IP network using multiple networks. The idea is to recover the intra-domain failure through inter-domain links with neighbors. Recently, Resilient Routing Reconfiguration (R3) is proposed as a novel routing protection scheme that guarantees performance and congestion-free [34]. These methodologies are efficient for intra-domain failures, but can not recover inter-domain traffic, nor large scale disasters.

Building dedicated backup infrastructures. Specially customized dedicated emergency response network (ERN) is introduced in [35] to sustain critical communications in emergencies. The idea is to use Advanced Metering Infrastructure (AMI), surveillance camera mesh networks, and enhanced cell phones for emergency communication when a disaster happens. ERN aims to solve large scale failures, however, building such dedicated backup infrastructures is very expensive and time-consuming

Using the existing Internet self-healing. Some approaches [10]–[12] attempt to explore and utilize the policy-compliant BGP path for the resilience enhancement. R-BGP is proposed in [12] to ensure that Internet domains stay connected as long as the underlying network is connected. R-BGP works by pre-computing a few strategically chosen failover paths. R-BGP provably guarantees that a domain will not become disconnected from any destination as long as it has a policy-compliant path to that destination after convergence. In fact, R-BGP focuses more on eliminating the packet loss caused by BGP dynamics and does not improve much on failure resilience. In MIRO [11], the authors present a multi-path inter-domain routing protocol that offers substantial flexibility, while giving transit domains control over the flow of traffic through their infrastructure and avoiding state explosion in disseminating reachability information. This work exploits the underlying path diversity and hence advances the availability of Internet routes under network failures. Path splicing [10] is a primitive for increasing reliability by composing routes from multiple routing protocol instances. The performance of path splicing under both intra-domain and inter-domain routing are evaluated to be promising when links fail. Although these mechanisms [10]–[12] are promising, their ability to Internet failures is inherently bounded by the BGP-based Internet routing structure as reported to be not resilient when various Internet emergencies happen [14].

VII. Conclusion

We are the first to evaluate the ability of two potential routing diversities, *i.e.*, peer relaxation and IXP participants reconnection, which are proposed to recover Internet failures beyond existing routing connectivities based on BGP policies. With most complete AS-level map and IXP dataset that can obtain, we have quantified the recovery ratio, path diversity and shifted paths in different failure scenarios. The evaluation results suggest that the two potential routing diversities pinpointed in this paper are promising means to recover the failed routings. As a first step towards the goal of utilizing these potential routing diversities, this paper has demonstrated the efficiency to adopt them. The next step about an efficient mechanism and system to turn the "potential" routings into "real" ones is described in our report [32].

Acknowledgment

We would like to thank Richard Yang for the insightful discussions with him at early stage of this work, and thank the anonymous reviewers for their valuable comments. This work is supported by NSF Award (CNS-0917233), NSFC (60873250, 60903182, 60921003, 61073171), 863 project (2007AA01Z216), fund for the Chinese doctoral program of higher education (20060003058), Tsinghua Research Plan Z02-2 and the Fundamental Research Funds for Central Universities. Any opinions, findings, and conclusions expressed are those of the authors and do not reflect the views of the funding sources.

References

[1] GENI Planning Group, "GENI: conceptual design, project execution plan," in GENI Design Document.

[2] NANOG, "North American Network Operators Group Mailing List Archive," http://www.merit.edu/mail.archives/nanog/.

[3] A. Ogielski and J. Cowie, "Internet routing behavior on 9/11 and in the following weeks," www.renesys.com/tech/presentations/pdf/renesys-030502-NRC-911.pdf.

[4] A. Popescu, T. Underwood, and E. Zmijewski, "Quaking tables: the Taiwan earthquakes and the Internet routing table," in *NANOG 39*, 2007.

[5] N. M. L. Archive, "Fiber Cut in SF Area," http://mailman.nanog.org/pipermail/nanog/2009-April/thread.html#start.

[6] S. Wilcox, "Quaking Tables: The Taiwan Earthquakes and the Internet Routing Table," http://www.thedogsbollocks.co.uk/tech/0705quakes/AMSIXMay07-Quakes.ppt, 2007.

[7] F. Wang, Z. M. Mao, J. Wang, L. Gao, and R. Bush, "A measurement study on the impact of routing events on end-to-end internet path performance," in *Proc. 2006 SIGCOMM*, pp. 375–386.

[8] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "Medusa—new model of internet topology using k-shell decomposition," http://www.mendeley.com/research/medusa-new-model-of-internet-topology-using-kshell-decomposition, 2006.

[9] W. Mühlbauer, S. Uhlig, A. Feldmann, O. Maennel, B. Quoitin, and B. Fu, "Impact of routing parameters on route diversity and path inflation," *Computer Network*, vol. 54, pp. 2506–2518, Oct. 2010.

[10] M. Motiwala, M. Elmore, N. Feamster, and S. Vempala, "Path splicing," in *Proc. 2008 ACM SIGCOMM*, pp. 27–38.

[11] W. Xu and J. Rexford, "MIRO: multi-path interdomain routing," in *Proc. 2006 ACM SIGCOMM*, pp. 171–182.

[12] N. Kushman, S. Kandula, D. Katabi, and B. M. Maggs, "R-BGP: staying connected in a connected world," in *2007 USENIX/SIGCOMM NSDI*.

[13] S. Secci, H. Ma, B. Helvik, and J.-L. Rougier, "Resilient inter-carrier traffic engineering for internet peering interconnections," *IEEE Trans. Network and Service Management*, vol. 8, no. 4, pp. 274–284, Dec. 2011.

[14] J. Wu, Y. Zhang, Z. M. Mao, and K. G. Shin, "Internet routing resilience to failures: analysis and implications," in *2007 CoNEXT*.

[15] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, "In search of the elusive ground truth: the Internet's AS-level connectivity structure," in *2008 ACM SIGMETRICS*.

[16] C. Hu, K. Chen, Y. Chen, and B. Liu, "Evaluating potential routing diversity for Internet failure recovery," in *Proc. 2010 INFOCOM*, pp. 1–5.

[17] Y. R. Yang, H. Xie, H. Wang, L. E. Li, Y. Liu, A. Silberschatz, and A. Krishnamurthy, "Stable route selection for interdomain traffic engineering," *IEEE Network*, pp. 90–97, 2005.

[18] L. Gao, "On inferring autonomous system relationships in the Internet," *IEEE/ACM Trans. Networking*, vol. 9, no. 6, pp. 733–745, Dec. 2001.

[19] A. links, "UCLA IRL," http://irl.cs.ucla.edu/topology/.

[20] IRR, "Internet Routing Register," http://www.irr.net.

[21] ROUTEVIEWS, "Routeviews Project," http://www.routeviews.org/.

[22] RIPE, "Routing Information Service," http://www.ripe.net/projects/ris/.

[23] R. Oliveira, B. Zhang, and L. Zhang, "Observing the evolution of Internet AS topology," in *2007 ACM SIGCOMM*.

[24] K. Chen, D. Choffnes, R. Potharaju, Y. Chen, F. E. Bustamante, D. Pei, and Y. Zhao, "Where the sidewalk ends: extending the Internet as graph using traceroutes from p2p users," in *2009 CoNEXT*.

[25] Packet Clearing House, "Internet exchange directory," http://www.pch.net/resources/data.php?dir=/exchange-points.

[26] Peeringdb, "Peering database," http://www.peeringdb.com.

[27] Euro-IX, "Europe's leading internet exchange points," http://www.euro-ix.net/.

[28] J. Xia and L. Gao, "On the evaluation of AS relationship inferences," in *2004 IEEE GLOBECOM*.

[29] Y. He, G. Siganos, M. Faloutsos, and S. V. Krishnamurthy, "A systematic framework for unearthing the missing links: measurements and impact," in *2007 USENIX/SIGCOMM NSDI*.

[30] Networkworld, "ISP spat blacks out net connections," http://www.networkworld.com.

[31] R. Bush, O. Maennel, M. Roughan, and S. Uhlig, "Internet optometry: assessing the broken glasses in Internet reachability," in *2009 IMC*.

[32] K. Chen, C. Hu, X. Wen, Y. Chen, and B. Liu, "Towards Internet Emergency Response via Reconfiguration in Internet eXchange Points (Tech. Report)," http://nskeylab.xjtu.edu.cn/people/huc/Pub/IER_report.pdf, 2011.

[33] H. Wang, Y. R. Yang, P. H. Liu, J. Wang, A. Gerber, and A. Greenberg, "Reliability as an interdomain service," in *2007 SIGCOMM'07*.

[34] Y. Wang, H. Wang, A. Mahimkar, R. Alimi, Y. Zhang, L. Qiu, and Y. R. Yang, "R3: resilient routing reconfiguration," in *2010 SIGCOMM*, pp. 291–302. Available: http://doi.acm.org/10.1145/1851182.1851218

[35] M. LeMay and C. A. Gunter, "Supporting emergency-response by retasking network infrastructures," in *2007 HotNets-VI*.

Chengchen Hu received his Ph.D. degree from the department of computer science and technology of Tsinghua University in 2008. He worked as an assistant research professor in Tsinghua University from Jun. 2008 to Dec. 2010 and is currently an associate professor in the Department of Computer Science and Technology of Xi'an Jiaotong University. His main research interests include computer networking systems, network measurement and monitoring.

Kai Chen is an Assistant Professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. He received his Ph.D. degree in Computer Science from the Northwestern University. He is particularly interested in finding simple yet deep and elegant solutions to real networking and system problems.

Yan Chen is an Associate Professor in the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL. He got his Ph.D. in Computer Science at University of California at Berkeley in 2003. His research interests include network security, and measurement and diagnosis for large scale networks and distributed systems. He won the Department of Energy (DoE) Early CAREER award in 2005, the Department of Defense (DoD) Young Investigator Award in 2007, and the Microsoft Trustworthy Computing Awards in 2004 and 2005 with his colleagues. Based on Google Scholar, his papers have been cited for over 2,600 times.

Bin Liu was born in 1964. He is now a Full Professor in the Department of Computer Science and Technology, Tsinghua University. His current research areas include high performance switches/routers, network processors, high speed security and greening the Internet. Bin Liu has received numerous awards from China including the Distinguished Young Scholar of China and won the inaugural Applied Network Research Prize sponsored by ISOC and IRTF in 2011.

Athanasios V.Vasilakos is currently Visiting Professor National Technical University of Athens (NTUA), Athens, Greece. He served or is serving as an Editor for many technical journals, such as the IEEE TNSM, IEEE TSMCPART B, IEEE TITB, ACM TAAS, the IEEE JSAC special issues of May 2009,Jan 2011,March 2011. He is Chairman of the Council of Computing of the European Alliances for Innovation.