Computer Networks 133 (2018) 183–194

Contents lists available at ScienceDirect

Computer Networks

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

A study on geographic properties of internet routing

Jessie Hui Wang*, Changqing An

Institute for Network Sciences and Cyberspace, Tsinghua University, China

ARTICLE INFO

Article history:
Received 15 March 2017
Revised 22 July 2017
Accepted 17 January 2018

Keywords:
Internet
Routing
Geography
Circuitousness
Centrality

ABSTRACT

In order to make a better control over the routing of Internet traffic, more and more researchers and governments want to understand how international reachability depends on individual countries. It has been necessary and valuable for us to study the geographic properties of Internet routing. In this paper, we conduct a measurement study on the dataset from 2011 to 2015 to understand two geographic properties of Internet routing: *geographically routing circuitousness* of paths and *geographically routing centrality* of countries and continents. Our analysis shows that the routing circuitousness of our Internet is deteriorating in these years. We also find that United States, Great British, France and Germany have most control over the data transfer in the Internet, but their farness centrality indexes are not smallest. Furthermore, our temporal analysis on the routing dependence among countries and continents finds out the importance of Europe was decreasing comparing with its competitor North America in the past years.

1. Introduction

The routing in the Internet is determined by the technical and business considerations of Internet Service Providers (ISPs). Roughly speaking, current intra-domain routing based on Open Shortest Path First (OSPF) routing protocol tries to minimize congestion on all intra-domain links, while current inter-domain routing based on Border Gateway Protocol (BGP) tries to provide a way for ISPs to enforce their business agreements. Both routing protocols do not take any geographical factors into consideration, which raises concerns on both network security of nations and efficiency of networking resource consumption.

Historically the Internet is regarded as a virtual world built on logical address of endpoints, i.e., IP address. Researchers only pay attention to network paths traversed by data packets. Therefore, previous efforts of researchers usually focused on the Internet's network layer topology and tried to answer questions such as "which ISPs are most important?" and "which routers are most important?"

However, as the Internet is growing to be more and more influential in our daily life, government control over the treatment of Internet traffic becomes more common, and many people will want to understand how international reachability depends on individual countries and to adopt strategies either for enhancing or weakening the dependence on some countries [1]. People have proposed the concepts of *national routing, Boomerang Routing, Internet sovereignty*, etc. [2–4]. The basic argument is that it is potentially insecure as traffic flows between two countries going through a third country, and one country may want to avoid its traffic flows going through some other countries if not necessary. Hence, it has been necessary and valuable for us to study the geographic properties of Internet routing.

In this paper, we conduct a measurement study to understand two geographic properties of Internet routing: *geographically routing circuitousness* of paths and *geographically routing centrality* of countries and continents.

Circuitousness means a traffic flow goes through a much longer geographical distance than the geographical distance from its source directly to its destination. Intuitively, it is not a uncommon phenomenon because Internet routing is trying to find paths with better performance under the constraint of business agreements. It does not take geographical distance into consideration. The resulting circuitous path may be "best" from the viewpoint of network load and congestion. But circuitousness can often be an indicative of a routing problem which deserves more careful examination [5]. A circuitous path may increase the risk of being wiretapped. Furthermore, circuitousness also suggests traffic flows are consuming more network resources than necessary, and it might be possible for us to reduce their resource consumption by improving network planning to avoid circuitousness, e.g. increasing capacity of some links or establishing new links [6].

In the first part of this paper, we calculate circuitousness ratios for paths collected during the period from 2011 to 2015. Our study shows the routing circuitousness of our Internet is deteriorating in

* Corresponding author.
E-mail address: hwang@cernet.edu.cn (J.H. Wang).

https://doi.org/10.1016/j.comnet.2018.01.032
1389-1286/

these years. Particularly, we group these paths according to various features, such as the number of Autonomous Systems (ASs) on the path, the number of continents, the number of countries, and geographical regions of the path. We then study the circuitousness distribution of each group in these years. Our measurement shows that statistically the circuitousness is increasing for most groups.

The second part of this paper focuses on centrality analysis. As far as we know, previous works usually focus on network layer topology of the Internet, and there is very few work on geographical topology of the Internet. In this paper, we study geographic paths traversed by data packets, and try to answer the questions such as "which countries are most important for the routing in the Internet?", or "which countries are at the center of the Internet?". Intuitively, it should be United States. But how to evaluate? How much difference between the first and the second most important country? Was there any changes in the past years? For a particular country, which countries are most important for its Internet traffic flows?

We construct topologies of the Internet at continent level and country level, and define several centrality indexes to evaluate the importance of countries and continents. Based on them, we identify important continents and countries from our geographical layer map of the Internet. We also define a metric to evaluate the routing dependence of one country on the other country. Our analysis shows that United States, Great British, France and Germany are most important countries for the transit of Internet traffic flows. In other words, these four countries have most control over the data transfer in the Internet. But their farness centrality indexes, i.e., average distances to other countries, are not smallest. Our temporal analysis also finds out the importance of Europe was decreasing comparing with its competitor North America. Most countries increasingly depend on United States to transfer their data flows, while Russia continuously depend on Great British more than United States and its dependence on United States was continuously decreasing in the past years.

Our paper is organized as follows. In Section 2, we present an overview of prior related works. Section 3 introduces the data sets we exploit and how we prepare them. In Section 4 we report our observations on the circuitousness of the Internet routing. Particularly, we study the paths with different lengths and in different regions. Section 5 presents a study on the geographical topology of Internet, listing important countries and continents. We also present the routing dependence among different countries. We conclude our paper in Section 6.

2. Related work

In 2002, Subramanian et al. have conducted measurement and analysis on geographic properties of Internet routing [5]. They propose to consider the geographic path traversed by packets, not just the network path. The circuitousness of Internet routes is one of the geographical properties they studied in the paper. It has been more than 15years after they conducted their measurements. In this paper, we exploit data sets from 2011 to 2015, showing the changes of circuitousness in these years. We also compare the properties of recent Internet with 15years ago when possible, and study the circuitousness of Internet traces with different lengths and in different regions.

In 2012, Matray et al. report their works on spatial properties of Internet routes in [7]. In the paper, motivated by the argument that the geographic layout of the physical Internet inherently determines important network properties and traffic characteristics, they conduct a geographically dispersed traceroute campaign, and embed the extracted topology into the geographic space by applying a novel IP geolocalization service, called Spotter. The investigations presented in the paper include the length distribution of Internet links, and also a brief study on the circuitousness and asymmetry of end-to-end Internet routes.

Our previous work on routing circuitousness focuses on intercontinental traffic flows [6]. In the paper, we report several intercontinent cases with large circuitousness, and investigate possible causes for their circuitousnesses based on multiple information sources such as PeerDB. Our study demonstrates the possibility of mitigating circuitousness by careful network planning.

As far as we know, there is only one paper on country path analysis, which is published by Karlin et al. in 2009 [1]. The authors point out that as government control over the treatment of Internet traffic becomes more common, many people will want to understand how international reachability depends on individual countries and to adopt strategies either for enhancing or weakening the dependence on some countries. They conduct analysis based on betweenness centrality, and present top countries with largest betweenness centrality. In this paper, we use a different method and dataset to derive country level paths, and our result is consistent with theirs on the top four countries, which enhances the credibility of results. Besides betweenness centrality, we also define more metrics, such as farness centrality, degree centrality and routing dependence, to evaluate the importance of countries. We also present a study on temporal variation of these metrics in these years, and find that the importance of some countries and continents is decreasing.

In order to study geographical properties of Internet routing, we must be able to determine geographical locations of endpoints in the Internet, i.e., mapping each IP address to its geographical location. This research area, which is called as *geolocation*, has drawn a lot of attentions in both academia and industry. Researchers have proposed a lot of algorithms to improve the accuracy and precision of geolocation [8–15]. The result of geolocation has been applied in many areas, such as network security and online advertisements. And a lot of companies or organization have published their geolocation databases, as a paid service [16–19], or a free service [20–23]. But in [24], the authors investigate several databases, and find that these databases work well at country level, but may not be consistent on a finer granularity than country. Therefore, it is still an open research area with great challenges.

3. Datasets and data preparation

Our analysis in this paper is based on two public data sets, i.e., CAIDA UCSD IPv4 /24 Routed Topology Dataset [25] and Maxmind GeoLite2 Dataset [26].

The CAIDA Dataset is consist of a lot of paths collected by a globally distributed set of Ark monitors. The monitors use team-probing to distribute the work of probing the destinations among the available monitors. Destinations are selected randomly from each routed IPv4 /24 prefix on the Internet such that a random address in each prefix is probed approximately every 48 h (one probing cycle). Monitors collect data by sending scamper probes continuously to destination IP addresses. Scamper is a successor of skitter, and it probes destinations with ICMP packets, using the Paris traceroute technique (ICMP-paris) to improve measurement integrity across load-balanced links. Data has been collected continuously since September 13, 2007. In this study, we use three snapshots, i.e., January of 2011, January of 2013 and January of 2015. Ark monitors are grouped into three probe teams, and each of our snapshot consists of one probe cycle of each probe team. Therefore each snapshot in fact covers a whole set of paths from ark monitors to all routed IPv4 /24 prefixes.

Each path probed by one monitor to one destination in CAIDA dataset is recorded as a sequence of IP addresses, and together with other information such as Round Trip Time (RTT) of both intermediate hops and the destination. In this paper, we call it "an

Table 1
The statistics of our data sets.

date	# paths	monitor	dst ip	src ctry	dst ctry	src cont	dst cont	src asn	dst asn
201101	673,167	46	672,092	26	190	6	6	45	10,246
201301	914,513	43	913,094	24	198	6	6	41	11,928
201501	523,017	56	522,501	27	193	6	6	53	11,678

IP level path" when we emphasize the sequence of IP addresses of the path.

We then use Maxmind to locate each IP address on this path, including its latitude, longitude, country, continent, and AS number. Now we can transform each *IP level* path into *country level* path, and *continent level* path, and *AS level* path. Since it is usual that multiple consecutive routers belong to a same AS, a same country or a same continent, the hop number of country level path, continent level path or AS level path is usually shorter than the hop number of IP level path. We also can calculate geographical distance between the source and destination and geographical distance of each hop using the information of latitude and longitude of IP addresses on the path. Based on these continent level paths, country level paths, and geographical distance information, we can construct geographical topology maps for the Internet, and conduct a study on properties of these maps.

One thing to note is that not every path collected by CAIDA can be used for our analysis. Some paths do not have complete information, i.e., some routers on the path do not respond to the monitor so that we do not have IP addresses of these hops. Some paths have complete sequences of IP addresses, but some of these IP addresses can not be mapped to their geographical locations because Maxmind does not provide their information. These paths with incomplete information are filtered out before our analysis. Table 1 presents the statistics of the final data sets we use in this paper.

4. Circuitousness and the impact of path length and regions

From the latitude and longitude of IP addresses on a path, we can calculate geographical distance between the source and the destination, and also geographical distance of each hop. Similarly as [7], we define circuitousness ratio of one path $P = (r_1, r_2, \ldots, r_n)$ as follows:

$$R(P) = \frac{\sum_{i=1\ldots n-1} D(r_i, r_{i+1})}{D(r_1, r_n)}, \quad (1)$$

wherein n is the number of IP addresses on this path, r_i is the ith IP address on the path, and $D(r_i, r_j)$ is the geographical distance, i.e., great circle distance, between r_i and r_j. $R(P)$ takes the value of 1 in the theoretical case when the links of the path exactly follow the great circle course between the source and the destination. In practical cases the value of R is greater than 1 and its magnitude reflects the extent of the path's deviation from the ideal course.

We calculate circuitousness ratio for each path in our data set, and plot its distribution of each snapshot in Fig. 1. It shows the Internet has more and more severe circuitousness from 2011 to 2015. The percentage of paths with $R(P) < 2$ decreased from 0.765 in 2011 to 0.661 in 2015, and the percentage of paths with $R(P) < 3$ decreased from 0.871 in 2011 to 0.826 in 2015.

To help us understand the situation, we try to find more historical measurement results from previous research efforts. In [5], the authors presented their results based on two small data sets. The Paxson data set in 1995 was gathered in late 1995, early in the life of the commercial Internet, and includes traceroutes conducted amongst the 33 nodes (mainly at academic locations). The 2000 data set is collected at academic sites connected by Internet2 backbone, a home cable modem network, and Microsoft Research

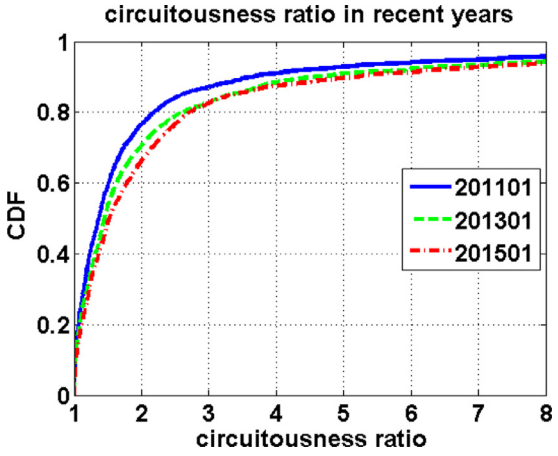

Fig. 1. Circuitousness of internet routing in recent years.

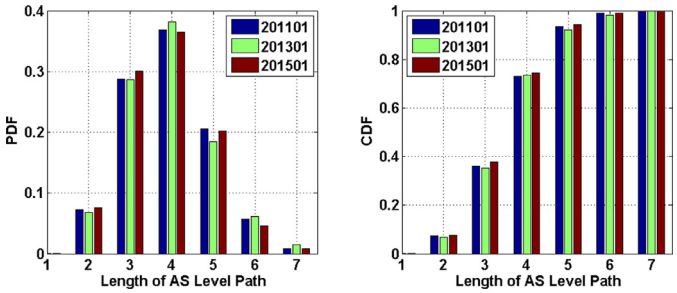

Fig. 2. Distribution functions of AS length in recent years.

network. According to their measurement, the percentage of paths with $R(P) < 2$ is about 0.65 in 1995 and 0.70 in 2000, the percentage of paths with $R(P) < 3$ is about 0.78 in 1995 and about 0.85 in 2000. The authors concluded the circuitousness was improving from 1995 to 2000 because the Internet was more and more richly connected.

Although these two data sets contain far fewer paths than our data set, it provides an interesting data point for comparison. Roughly speaking, the circuitousness of paths in our 2015 snapshot has been worse than paths in their 2000 data set.

In this section, we would study the circuitousness from various aspects, and try to find whether the path length at AS level, continent level and country level affect circuitousness ratio of the path.

4.1. Circuitousness of paths with different AS length

Let us define the number of ASs traversed by a path as its *AS length*. We compute the distribution of AS length of paths collected in the years of 2011, 2013 and 2015, and plot the results in Fig. 2. Since the data set is collected by a same algorithm continuously, it makes sense to compare the distribution of AS level path length in different years. As shown in the figure, most of paths traverse four different ASs in the data set. We can see that there is only small change in the distribution during recent years.

Now we try to study if there is any relationship between circuitousness and AS path length. We classify paths of each snapshot

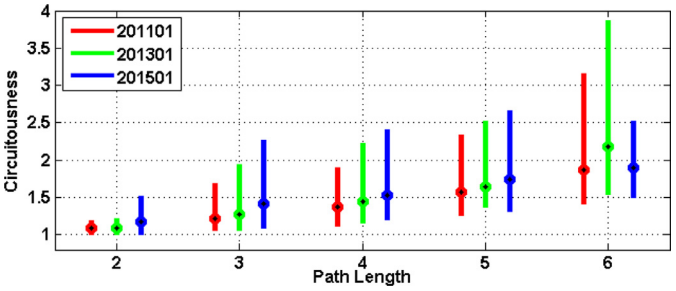

Fig. 3. AS length and circuitousness ratio.

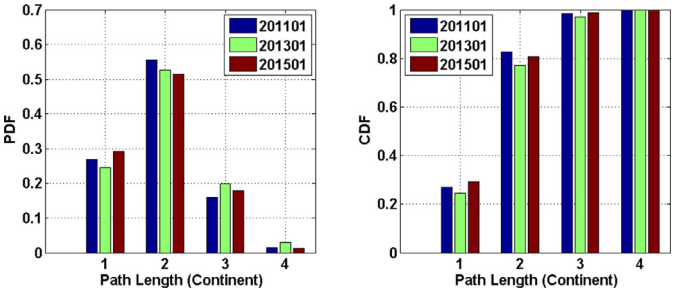

Fig. 4. Distribution of continent path length.

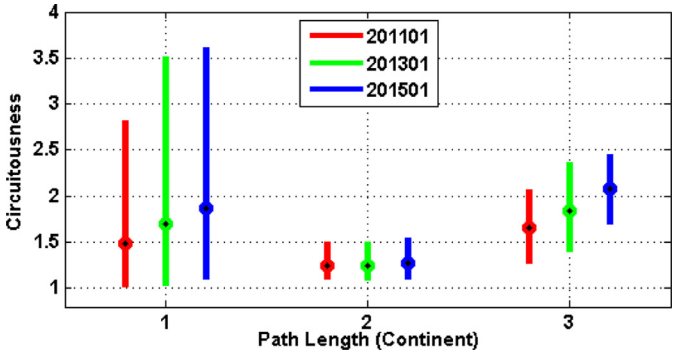

Fig. 5. Continent path length and circuitousness ratio.

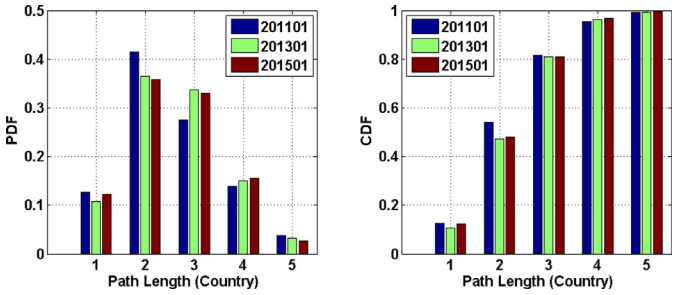

Fig. 6. Distribution of country path length.

into different sets according to their lengths. For each set, we compute the 25th, 50th and 75th percentiles of circuitousness ratios of paths in the set. Then we make a box plot to show our results in Fig. 3. In this figure, each vertical line represents paths in one set, i.e., with a particular AS path length and in a particular snapshot. The upmost point of the vertical line marks the 75th percentile of circuitousness ratio of paths in the set, the bottom point of the line marks the 25th percentile, and the point on the line marks the 50th percentile, i.e., the median value.

Here, we only consider paths with AS length $x \in [2, 6]$, since the number of paths with $x < 2$ or $x > 6$ is very small, as shown in Fig. 2, which means these sets are less important and statistics of these sets are prone to be affected by measurement bias.

Fig. 3 shows a clear trend of increasing from 2011 to 2015, with only very few exceptions, e.g., $x = 6$ in 2015. We can compare the results from two different aspects, i.e., circuitousness of paths with different path lengths in a same year, and circuitousness of paths in different years with a same path length. Then we have the following two observations.

- **Observation 1:** *Statistically, circuitousness of paths with a larger AS length is more severe than paths with a smaller AS length.*
- **Observation 2:** *Statistically, circuitousness is becoming more and more severe in these years for most sets of paths with a same AS length.*

4.2. Circuitousness of paths with different continent length

Fig. 4 shows the distribution of path length at continent level, i.e., the number of continents appeared on one path. We can see that in the data set more than a half of paths traverse two continents, and almost all paths traverse less than or equal to three continents. The changes of distribution is very small during these years, which is possibly because the data is collected by a same algorithm.

We also look into paths whose continent path length is 4. Almost all these paths are going through these continent sequences: *SA|NA|EU|AS*, *AF|EU|NA|AS*, and *OC|NA|EU|AS*.

Now we classify paths of each snapshot into different sets according to their continent length. We plot the 25th, 50th and 75th percentiles of the circuitousness ratio of paths in these sets as Fig. 5.

Compared with paths that traverse two or more continents, intra-continent paths tend to have a much larger circuitousness ratio. One important reason is that the geographical distance between source and destination is often smaller for intra-continental paths than inter-continental paths, which is favourable for larger circuitousness ratio.

However, we can also find that the circuitousness of intra-continental paths are more and more severe in these years, which can not be explained by the above argument. The deterioration trend also appears for the set of paths that traverse three continents, which might imply that some inter-continental traffic flows are exchanged at a less appropriate third continent, or at less appropriate locations of the third continent, e.g., further inland locations.

The set of paths with a continent length of 2 presents a bit better message. Its three statistics roughly keep stable during these years.

We summarize our observations as follows:

- **Observation 3:** *Statistically, circuitousness is becoming more and more severe during these years for intra-continental paths and inter-continental paths that traverse more than two continents.*
- **Observation 4:** *Statistically, circuitousness roughly keeps stable for intercontinental paths that go directly from source continents to destination continents.*

4.3. Circuitousness of paths with different country length

Fig. 6 shows the distribution of country level path length, i.e., the number of countries appeared on one path. We can see that there is almost no path longer than 5 countries. The proportions of paths with different lengths changed slightly. Although the trend is not evident, it seems there are more paths with length longer than 2 countries in 2013 and 2015.

Similarly as AS path length and continent path length, we also make a box plot to study the relationship between circuitousness

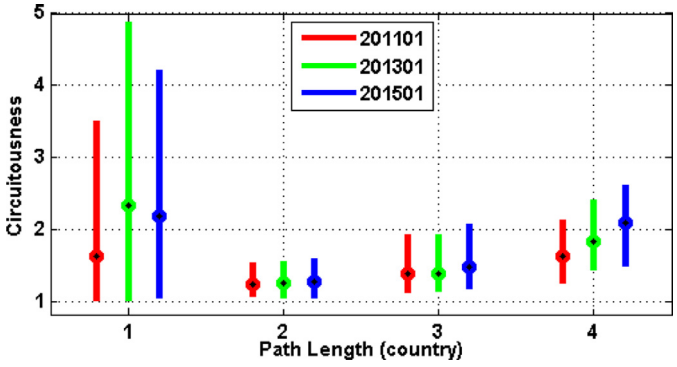

Fig. 7. Country path length and circuitousness ratio.

and country path length in Fig. 7. We summarize our observations from Fig. 7 as follows:

- **Observation 5:** *Statistically, intra-country paths tend to have a largest circuitousness ratio. For inter-country paths, circuitousness is worse for paths that traverse more countries.*
- **Observation 6:** *Statistically, circuitousness deteriorates for all paths that traverse three or more countries, while it keeps relatively stable for paths that go directly from source countries to destination countries.*

4.4. Impact of tier-1 ASs in paths

As we know, tier-1 ASs have their own global backbone networks, and play a very important role in the Internet routing. Therefore, we would like to have a look at their routing behaviors and their impacts on routing circuitousness.

Although "tier-1 AS" is a well-defined concept, it is not easy to list tier-1 ASs concretely, because it is difficult to learn the details of peering agreements of ASs. In our study, we consider eighteen ASs identified by CAIDA measurement project as tier-1 ASs.[1]

Fig. 8 presents statistics of the number of tier-1 ASs in paths. Please note theoretically it is impossible to see three tier-1 ISPs in one path according to the definition of tier-1 AS. Our measurement is consistent with this assert, i.e., there is nearly zero paths that traverse three or more tier-1 ASes. Fig. 8 also shows that the proportion of paths that traverse two tier-1 ASs is stable in recent years.

In Fig. 9, we plot circuitousness ratios of paths with different number of tier-1 ASs in 2011, 2013 and 2015. It is safe for us to summarize the following observations:

- **Observation 7:** *Statistically, paths with 2 tier-1 ASs tend to be more circuitous than paths with no tier-1 AS or only one tier-1 AS.*
- **Observation 8:** *Statistically, circuitousness does not become better in these years, except only one exception, i.e., circuitousness of paths with 2 tier-1 ASs was improved from 2013 to 2015.*

By definition, when there are two tier-1 ASs in a path, these two ASs must be connecting directly under a peer-to-peer agreement, which is called as "valley-free" rule. When two ASs peer with each other, there can be two kinds of routing policies: hot-potato routing or cold-potato routing. In hot-potato routing, an ISP hands off traffic to a downstream ISP as quickly as it can. Cold-potato routing is the opposite of hot-potato routing where an ISP

[1] In this paper, we regard these ASs as tier-1 ASs: AS7018, AS209, AS174, AS3320, AS3257, AS286, AS3356, AS6830, AS2914, AS5511, AS1239, AS6453, AS6762, AS12956, AS1299, AS701, AS2828, and AS6461.

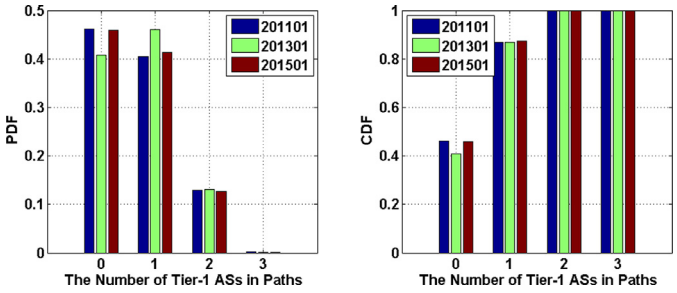

Fig. 8. The number of tier-1 ASs in paths.

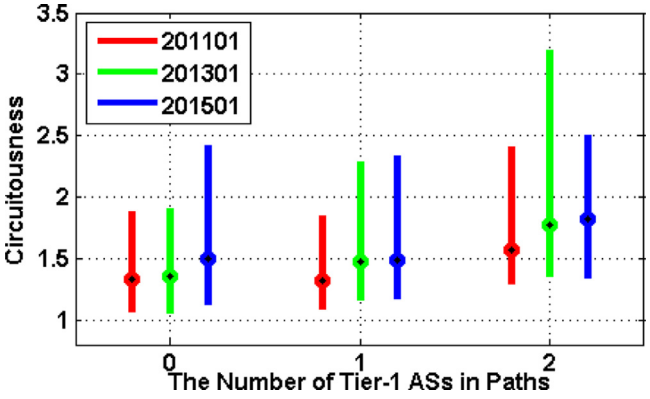

Fig. 9. The number of tier-1 ASs in paths and circuitousness ratio.

carries traffic as far as possible on its own network before handing it off to a downstream ISP. In [5], based on the 2000 data set, the authors find that the first AS tends to hand off traffic quickly to the second AS who carries it for a much greater distance, which means hot-potato routing policy is more popular.

We conduct a similar analysis on the data set of recent years. We compute the fraction of the end-to-end linearized distance that is accounted for by each individual tier-1 AS and plot our result in Fig. 10. There are four cdf curves, which are respectively the fraction accounted for by the tier-1 AS in paths with only one tier-1 AS, the fraction accounted for by the first tier-1 AS in paths with two tier-1 ASs, the fraction accounted for by the second tier-1 AS in paths with two tier-1 ASs, and the fraction accounted by two tier-1 ASs in total in paths with two tier-1 ASs.

We can see there is not much difference between the first and second tier-1 AS in paths with two tier-1 ASs in 2013, while in 2015, the fraction of the first tier-1 AS tends to be slightly larger than the second, which is a possible hint of cold potato routing. This is opposite to the result in the previous work.

Fig. 10 in fact shows a comparison of statistical results, in which we do not compare the distance of the first tier-1 AS and the second of a same path directly. We also try to directly investigate whether the distance of the first tier-1 AS is shorter or longer than the second. Our investigation result shows that in 52.4% of paths the first transmits a longer distance in 2013, while in 2015 the percentages of paths where the first is longer is 54.2%. It also suggests cold potato routing is a little more popular.

- **Observation 9:** *When one tier-1 AS is connecting directly with the other tier-1 AS, cold potato routing is used a little more frequently.*

Fig. 10 is a study on routing behavior across tier-1 ASs. We also study routing behaviors within a single tier-1 AS in Figs. 11 and 12.

Fig. 11 presents the number of non-zero-distance hops within one tier-1 AS on one path. Particularly, we remove the routing hops whose source points are very close to destination points, because

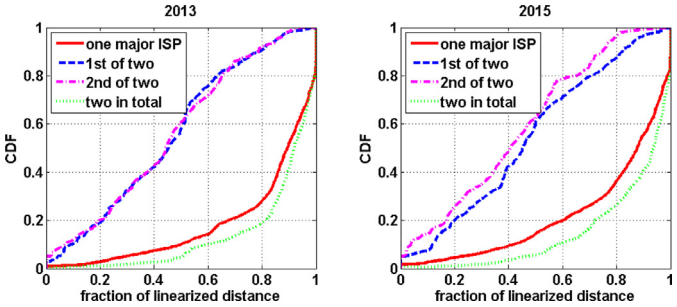

Fig. 10. Fraction of linearized distance accounted by tier-1 ASs.

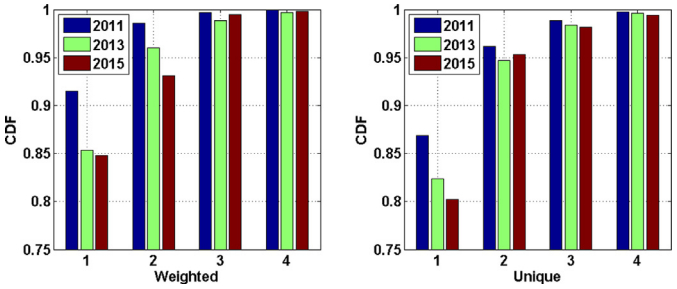

Fig. 11. Routing hops within tier-1 ASs.

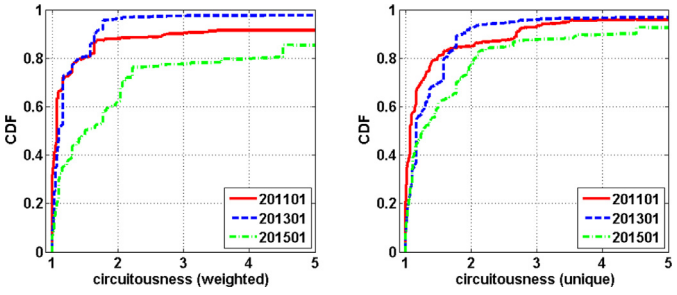

Fig. 12. Circuitousness of tier-1 ASs (>1000 km).

we think these zero-distance routing hops usually appear for purposes unrelated to traffic transmission.

Some intra-tier-1-AS segments may appear on different paths. For example, there are two paths in our data set, say (s_1, d_1) and (s_2, d_2). Both paths have a same segment in one tier-1 AS, say (m_{in}, m_{out}). In the left part of Fig. 11, we count this intra-tier-1-AS segment twice, and call it as "weighted cdf". In the right part, we count this segment only once, and it is called "unique cdf". Both plots show that the number of routing hops is increasing statistically, which implies the routing within tier-1 ASs might be more and more complex.

Tier-1 ASs are often used for long-distance traffic transmission. We focus on segments whose geographical distance between incoming points and outgoing points, e.g. m_{in} and m_{out}, are more than 1000km, and plot circuitousness ratio distribution of these segments in Fig. 12. Both the weighted cdf and unique cdf show that the circuitousness is worst in 2015.

We summarize our observations from Figs. 11 and 12 as follows.

- **Observation 10:** *For intra-tier-1-AS segments, the proportion with only one routing hop is decreasing, which implies routing within tier-1 ASs might be more and more complex.*
- **Observation 11:** *For intra-tier-1-AS segments whose routing hops are more than one and geographical distance is more than 1000*km, circuitousness is worse in 2015 than both 2011 and 2013.

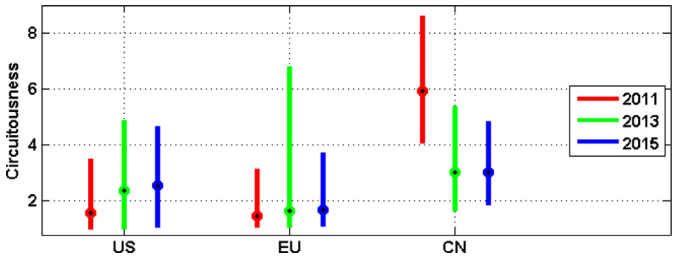

Fig. 13. Routing within regions: US, EU, and CN.

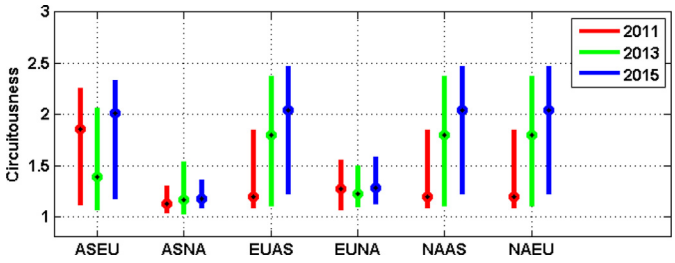

Fig. 14. Circuitousness of inter-continent paths.

4.5. Routing circuitousness of intra-region paths

Obviously it is unfair to compare intra-country routing circuitousness of different countries directly, since countries have different geographical sizes which may have important impacts on their circuitousness ratios. Therefore, we select three regions with comparable size for our study, i.e., United States, Europe, and China.

The result is presented in Fig. 13.[2] Roughly speaking, routing within Europe is the best, and routing within United States is better than China. Comparing circuitousness ratios in different years, we can see that China was making a big improvement, and routing within United States seems to have a slightly larger circuitousness. Routing circuitousness within Europe was relatively stable, but in 2013 the 75th percentile is much larger than other years, which means the worst 25% paths are more circuitous.

We can compare these results with Fig. 6 in [5], which studies routing circuitousness within Europe and United States in the year of 2000. By visual inspection, we can conclude the 50th and 75th percentiles of circuitousness ratio in recent years are larger than the year of 2000 for both Europe and United States. We do not compare 25th percentiles because they are very close to 1 in both two figures.

4.6. Routing circuitousness of inter-continental paths

For our study on paths across continents, we focus on three major continents, i.e., Europe, North America and Asia. Since there are less monitors and less Internet users [27] in other minor continents, we have a small number of paths across those continents in our data set. In order to avoid the influence of measurement bias, we do not calculate statistics for minor continents in this paper.

Our result is presented in Fig. 14. We can see that statistically the paths from Asia to North America, and paths from Europe to North America have smallest circuitousness, and they keep relatively stable performance in these years.

Among the other four source destination pairs, three pairs, i.e., from EU to AS, from NA to AS, and from NA to EU, were having

[2] Since there is no intra-CN path collected in the 2011 snapshot, in this paper we use statistics of intra-CN paths collected in the year of 2012 instead.

more and more deteriorated performance in terms of routing circuitousness, and this trend is monotonous, which gives more credence to the hypothesis that circuitousness was increasing.

The circuitousness of paths from Asia to Europe fluctuated from 2011 to 2015, but it is also true that the most recent time point is the worst.

Another observation is that paths from North America to the other two continents are more circuitous than paths from these continents to North America.

5. Routing centrality: Identifying important nodes in the Internet

Technically, routing in the Internet is controlled by Internet Service Providers. Therefore, previous efforts of researchers usually focused on the Internet's network layer topology and tried to answer questions such as "which ISPs are the most important?" and "which routers are the most important?" However, as the Internet is growing to be more and more influential in our daily life, many governments feel that they have to be able to understand or even control country level paths of traffic flows.

In this section, we would focus on country level topology of the Internet, and try to answer the questions such as "which countries are most important for routing in the Internet?", or "which countries are at the center of the Internet?". Intuitively, it should be United States. But how to evaluate? How much difference between the first and the second most important country? Was there any changes in past years? For a particular country, which countries are most important?

In this section, we will calculate several centrality indexes for each country based on our geographical layer map of the Internet, and try to answer the above questions. A centrality index is usually given in terms of a real-valued function on the vertices of a graph, where the values produced are expected to provide a ranking which identifies the most important nodes. The word "importance" has a wide number of meanings, leading to many different definitions of centrality. Here, we exploit three centrality indexes: *betweenness centrality, farness centrality*, and *degree centrality*. They are used to characterize countries based on different definitions of "importance". We also define a metrics *routing dependence* index to evaluate the importance of one country for the transmission of cross-country traffic flows of the other country.

5.1. Betweenness centrality

Betweenness centrality quantifies the number of times a node acts as a bridge along paths between two other nodes. It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network by Freeman [28]. In a telecommunications network, a node with higher betweenness centrality would have more control over the network, because more information will pass through that node. It reflects the importance of one node on the transfer of all traffic flows in the Internet.

The betweenness of a vertex v, denoted by $C_B(v)$, in a graph $G:=(V, E)$ with V vertices is computed as follows:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

wherein σ_{st} is the total number of paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

The above betweenness centrality index of a node scales with the number of pairs of nodes. Therefore it is infeasible to compare the indexes of nodes in graphs with different sizes. Since in our dataset different snapshots have different number of node pairs, in order to enable temporal analysis, we further conduct a normalization and calculate the normalized betweenness, $\overline{C_B}(v)$, as follows:

$$\overline{C_B}(v) = \frac{C_B(v) - \min_{\forall u \in V} C_B(u)}{\max_{\forall u \in V} C_B(u) - \min_{\forall u \in V} C_B(u)}. \tag{3}$$

Then the most important node, i.e. nodes with most control, would have a normalized betweenness of 1, while the normalized betweenness of the least important nodes is 0.

The other consideration is that all node pairs are treated equally in above equations, which implies the communication between any node pairs are equally important. However, some node pairs have more traffic flows to transfer, and the communication between them is more frequent and thus can be thought as more important. Let us denote the frequency of communications from node s to node d as w_{st}, which can be approximated by the number of IP paths from s to d, i.e., $w_{st} = \sigma_{st}$. Then we can define *weighted betweenness centrality*, denoted by $C_B^w(v)$, as follow:

$$C_B^w(v) = \sum_{s \neq v \neq t \in V} w_{st} * \frac{\sigma_{st}(v)}{\sigma_{st}} = \sum_{s \neq v \neq t \in V} \sigma_{st}(v) \tag{4}$$

To facilitate the comparison of different years, we also conduct a normalization on the weighted betweenness, which is similar as Eq. (3).

We first apply the above metrics on the study of importance of continents, and the results are plotted in Fig. 15. We can see that no matter weighted or not, the importance of Europe and Asia are both decreasing from 2011 to 2015, while the importance of North America does not change a lot. In 2011, the normalize betweenness of Europe and North America are all about 1, which implies they play approximately equally important roles in inter-continental traffic transmission if we view all continents equally. At the same time, in 2011, the weighted betweenness of NA is much more than Europe, implying that NA plays a more important role in data transmission between major continents.

The betweenness centrality index reflects the importance of one node on communication between all node pairs. We also look into each node pair, and list vectors (s, t, v) with highest $\frac{\sigma_{st}(v)}{\sigma_{st}}$ in Table 2. Take the first line as an example. It means 99.98% of traces from SA to AS depend on NA to complete their data transmission.

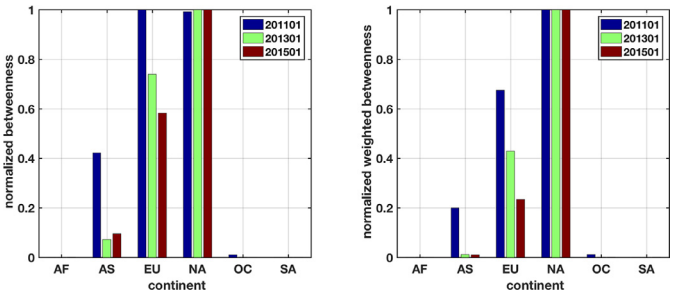

Fig. 15. Normalized betweenness of continents in recent years.

Table 2
(s, t, v) with $\frac{\sigma_{st}(v)}{\sigma_{st}} > 0.95$ and $\sigma_{st} > 1000$ in 2015.

s	t	v	$\frac{\sigma_{st}(v)}{\sigma_{st}}$	$\sigma_{st}(v)$	σ_{st}
SA	AS	NA	0.9998	6025	6026
AF	NA	EU	0.9980	2526	2531
AF	AS	NA	0.9930	1552	1563
AF	AS	EU	0.9898	1547	1563
OC	AS	NA	0.9849	2086	2118
SA	EU	NA	0.9820	8792	8953
OC	EU	NA	0.9585	5007	5224

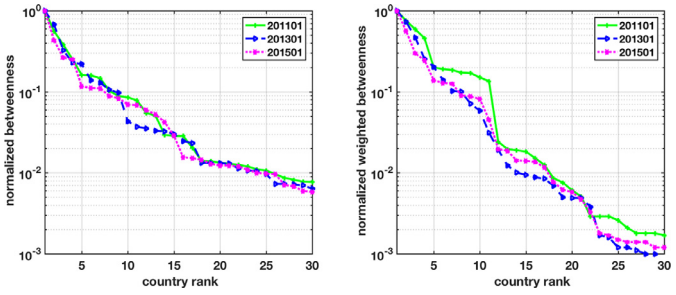

Fig. 16. Normalized betweenness of top countries in recent years.

Table 3
Normalized betweenness of top countries.

Rank	2011		2013		2015	
1	US	1.000	US	1.000	US	1.000
2	GB	0.559	GB	0.672	GB	0.432
3	FR	0.380	FR	0.326	FR	0.264
4	DE	0.251	DE	0.228	DE	0.252
5	NL	0.161	SE	0.220	NL	0.117
6	SE	0.159	NL	0.138	IT	0.111
7	IT	0.147	IT	0.129	SE	0.110
8	ES	0.103	ZA	0.106	ES	0.089
9	RU	0.088	ES	0.097	HK	0.083
10	HK	0.086	SG	0.043	IN	0.070

Table 4
Normalized weighted betweenness of top countries.

Rank	2011		2013		2015	
1	US	1.000	US	1.000	US	1.000
2	GB	0.753	GB	0.730	GB	0.565
3	FR	0.589	FR	0.468	FR	0.298
4	DE	0.459	DE	0.255	DE	0.245
5	NL	0.200	SE	0.202	NL	0.139
6	SE	0.190	NL	0.140	SE	0.128
7	ES	0.185	ES	0.102	HK	0.126
8	IT	0.172	ZA	0.100	AT	0.090
9	HK	0.170	IT	0.071	ES	0.088
10	NZ	0.150	NZ	0.058	IT	0.081

Considering the influence of measurement bias, we do not present the vectors with small σ_{st} even if they have large $\frac{\sigma_{st}(v)}{\sigma_{st}}$.

Fig. 16 presents the results of our study on betweenness of countries. We plot normalized betweenness and normalized weighted betweenness of top 30 countries. Obviously, the importance decreases very quickly as the rank increases, which shows the feature of "power-law". The betweenness of the 10th country has been less than 0.1. In other words, a few countries dominate the transfer of cross-country traffic flows. The slope of curves in the right figure is more steep than curves in the left figure, which suggests that top countries are more dominant for traffic flows between major countries, i.e., countries with more Internet users.

We then list top 10 countries with largest normalized betweenness in Table 3, and also list top countries in terms of normalized weighted betweenness in Table 4. In both tables, the top 4 countries are always United States, Great Britain, France and Germany. Furthermore, the betweenness centrality of United States is much larger than all other countries.

Similarly as Table 2, we also look into each pair of source country and destination country and list vectors (s, t, v) with $\frac{\sigma_{st}(v)}{\sigma_{st}} > 0.95$ and $\sigma_{st} > 1000$ in Table 5. To illustrate, let us take the first two lines as an example. It shows that all (100%) traffic flows from Australia and Brazil to China are transited by United States.

Table 5
(s, t, v) with $\frac{\sigma_{st}(v)}{\sigma_{st}} > 0.95$ and $\sigma_{st} > 1000$ in 2015.

s	t	v	$\frac{\sigma_{st}(v)}{\sigma_{st}}$	$\sigma_{st}(v)$	σ_{st}
AU	CN	US	1.0000	1880	1880
BR	CN	US	1.0000	3093	3093
ES	CN	GB	1.0000	2365	2365
FI	US	SE	1.0000	4042	4042
GB	CN	FR	1.0000	1301	1301
MU	US	ZA	1.0000	2487	2487
SG	CN	US	1.0000	4300	4300
SG	DE	US	1.0000	1410	1410
SG	KR	US	1.0000	1482	1482
ES	CN	US	0.9996	2364	2365
IE	KR	US	0.9993	1417	1418
BR	ES	US	0.9992	1235	1236
MU	US	GB	0.9980	2482	2487
BR	DE	US	0.9974	3053	3061
BG	CN	FR	0.9932	1319	1328
ES	US	GB	0.9919	4529	4566
CA	JP	US	0.9906	1259	1271
AU	DE	US	0.9818	2052	2090
DE	KR	US	0.9747	2043	2096

5.2. Farness centrality

In a connected graph, the more central a node is, the closer it is to all other nodes. Thus the sum of the length of the shortest paths between the node and all other nodes can be used to measure the centrality of one node in a network. In 1950, Bavelas defined a metrics called as *closeness centrality*, $C(t)$, to measure the centrality of one node $t \in V$ as follows:

$$C(t) = \frac{|V| - 1}{\sum_{s \neq t, s \in V} d(s, t)}, \tag{5}$$

wherein $|V|$ is the number of nodes in the graph, and $d(s, t)$ is the distance between node s and node t.

In this paper, we directly use the reciprocal of closeness $C(t)$, i.e., the average distance (hop count) of other nodes to the node under study to measure its centrality, because it has a clear physical meaning and is more intuitive for us to understand. Furthermore, as we know, in the Internet the hop count from s to t might not be equal to the hop count from t to s. Since the average hop count from one country to other countries is heavily affected by the deployment of monitors in the country, we then use the hop count from other countries to the country under study to measure its closeness centrality. Therefore, the metrics we use is defined as follows:

$$C_C(t) = \frac{\sum_{s \neq t, s \in V} d(s, t)}{|V| - 1}, \tag{6}$$

wherein $|V|$ is the number of nodes in the graph, and $d(s, t)$ is the distance from node s to node t. $C_C(t)$ is in fact "*farness*" instead of "closeness". In other words, the more central a node is, the smaller its farness is. Here $d(s, t)$, the distance from s to t, can be calculated in two different ways. There are a lot of IP level paths from the country s to the country d. These paths can be mapped to multiple unique country level paths. If each unique country level path is counted only once when calculating $d(s, t)$, we call the result of Eq. (6) as "farness". If each IP level path is counted once, i.e., the frequency of one unique country level path is equal to the number of IP level paths which can be mapped to this country level path, we call the result of Eq. (6) as "weighted farness", because the country level path is weighted by its frequency in our data set.

Fig. 17 shows the farness centrality of each continent. No matter weighted or not, the trend is clear that North America is more and more close to the center of the Internet while Europe is farther away. For Africa and Oceania, the trends of farness and weighted

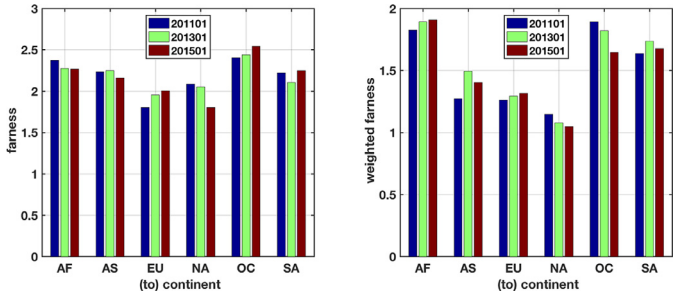

Fig. 17. Farness centrality of each continent.

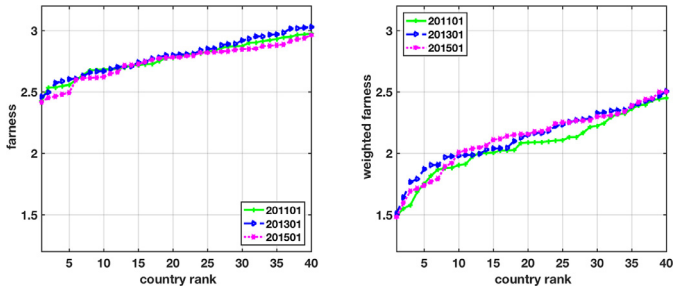

Fig. 18. Farness centrality of each country.

Table 6
Countries with smaller farness centrality.

Rank	2011		2013		2015	
1	FR	2.425	JP	2.460	PT	2.417
2	GB	2.536	FR	2.497	FR	2.453
3	TW	2.536	KR	2.572	KR	2.462
4	ZA	2.547	US	2.586	US	2.481
5	US	2.557	ZA	2.603	ES	2.494
6	CA	2.598	PT	2.607	GB	2.599
7	KR	2.637	TW	2.630	IN	2.611
8	FI	2.675	CY	2.657	JP	2.614
9	MX	2.678	FI	2.668	CA	2.618
10	HK	2.683	SE	2.669	FI	2.622

Table 7
Countries with smaller weighted farness centrality.

Rank	2011		2013		2015	
1	CN	1.490	US	1.508	US	1.482
2	US	1.551	ES	1.640	ES	1.595
3	ES	1.578	GB	1.766	FR	1.692
4	GB	1.687	CN	1.791	IN	1.714
5	NL	1.752	FR	1.871	GB	1.739
6	FR	1.818	JP	1.903	CN	1.770
7	CR	1.868	ph	1.905	CA	1.794
8	JP	1.878	IN	1.967	KR	1.893
9	CO	1.883	DE	1.976	DE	1.921
10	DE	1.903	NL	1.979	PT	2.009

farness are opposite. Take Africa as an example, its farness is decreasing and its weighted farness is increasing. The result shows there might be more shorter continent layer paths to Africa, but the routing decisions might be worse than before since most traffic flows to Africa are using longer paths.

Fig. 18 shows the farness centrality of top forty countries. Here we do not include the countries who have small number of traces in our dataset to avoid the influence of measurement bias. We can see that there is no much difference between different years. The farness increases linearly with the rank, while the weighted farness increases a little superlinearly.

From Table 6, we can see that France, Korea, United States and Finland are the only four countries that appear in the top lists of all three years. In terms of weighted farness, as shown in Table 7, United States, Spain, France, Great Britain, China, and Germany appear in the top lists of all three years. Obviously, the countries who are close to all other countries are not exactly same as the countries who play important roles on the transit of data flows.

Table 8
Countries with largest degre.

Rank	2011		2013		2015	
1	US	139	US	148	US	138
2	GB	93	GB	95	GB	93
3	FR	86	FR	81	DE	80
4	IT	65	NL	69	FR	71
5	DE	64	DE	66	NL	69
6	NL	51	IT	62	IT	66
7	RU	42	ES	42	ES	43
8	ES	41	SG	35	IN	34
9	SG	32	RU	32	SE	34
10	JP	28	SE	31	SG	31

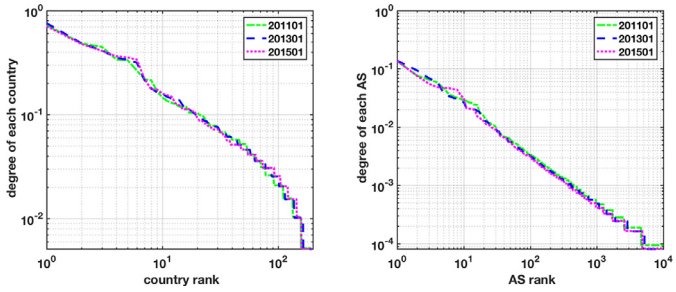

Fig. 19. Normalized degree vs. rank of countries and ASs.

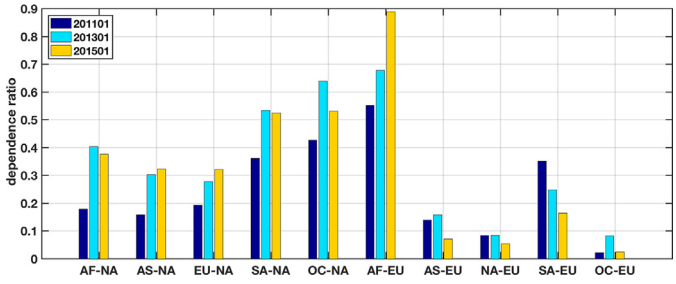

Fig. 20. Routing dependence of continents on EU and NA.

5.3. Degree centrality

Degree centrality is regarded as the first and conceptually simplest centrality metrics. In this study, we calculate the degree of each country and rank these countries according to their degrees. To allow comparison, we normalize the degree of one country as follows:

$$C_D(v) = \frac{|\ \{u|(u,v) \in E, u \in V\}\ |}{|V|} \tag{7}$$

wherein $|V|$ is the number of nodes in the graph G(V, E), i.e., the number of countries that appear in the snapshot. Obviously, $C_D(v) \leq 1$ for all v.

The top countries with largest degree centrality are presented in Table 8, where US, GB, FR and DE are the most important countries. It is roughly consistent with the lists of top countries with largest betweenness centrality shown in Tables 3 and 4.

Fig. 19 plots $C_D(v)$ of countries in recent years. For comparison, we also plot $C_D(v)$ of ASs in the same data set. Our study shows that the ranks of some countries fluctuate in these years. For example, during the period from 2011 to 2015, the rank of India is

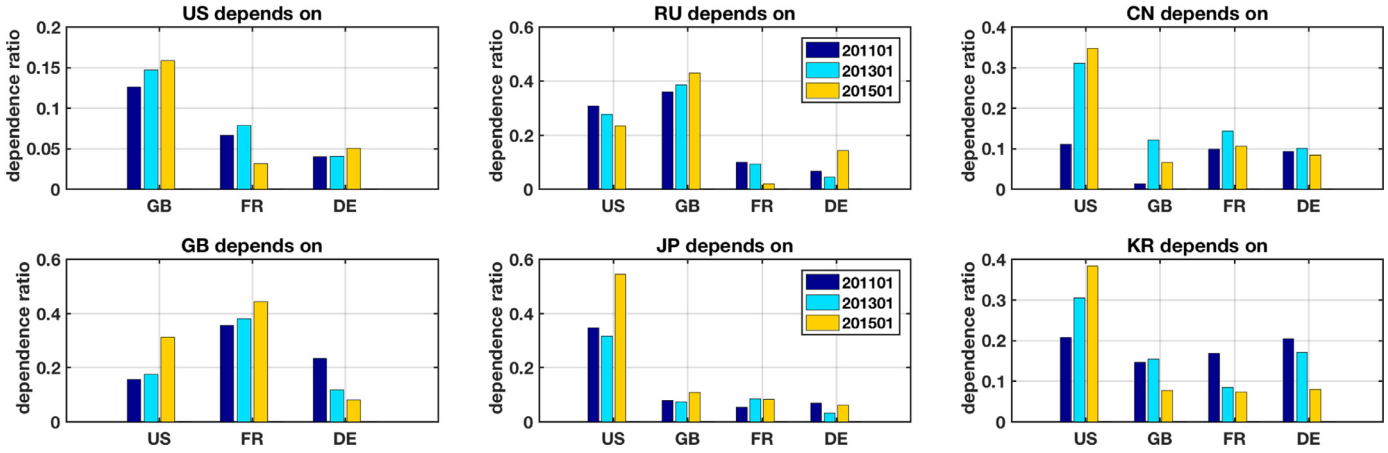

Fig. 21. Routing dependence on US, GB, FR and DE.

improved from 15 to 8, and Korea is from 22 to 16. But we can see that there is almost no change in the curves of normalized degree distribution after sorting countries and ASs. The degree distribution of sorted ASs is very close to a straight line, also the degree distribution of top 100 countries.

5.4. Routing dependency among continents and countries

In Table 5, we present some vectors (s, t, v) and their $\frac{\sigma_{st}(v)}{\sigma_{st}}$. To some extend, it represents how much the routing from s to t depends on v. In this subsection, we further define one metrics D_n^v, *routing dependence* of n on v, as the percentage of paths sourced from n or destined to n that go through v. Formally, we have:

$$D_n^v = \frac{\sum_{s \neq n}(\sigma_{sn}(v) + \sigma_{ns}(v))}{\sum_{s \neq n}(\sigma_{sn} + \sigma_{ns})}, \tag{8}$$

wherein σ_{sn} is the total number of paths from node s to node n and $\sigma_{sn}(v)$ is the number of those paths that pass through v.

As shown in Fig. 15, North America and Europe are most important continents on the transit of inter-continental traffic flows. Therefore we focus on these two continents and plot the dependence of other continents on them in Fig. 20. Roughly speaking, except Africa, all other continents rely on North America more than Europe, and the dependence of Asia and Europe on North America is increasing, while the dependence of North America and South America on Europe is decreasing.

Fig. 21 presents the dependence of some countries on the four countries with largest betweenness centrality. Almost all countries under study depend on United States more heavily in 2015 than 2011, with only one exception, i.e. Russia. In fact, traffic flows of Russia depend on Great British more than United States, and its dependence on Great British is increasing in these years. South Korea also has a clear trend, *i.e.*, it increasingly depends on United States, and decreasingly depends on all three countries in Europe.

6. Conclusion

As the Internet is growing to be more and more influential in our daily life, government control over the treatment of Internet traffic becomes more common. Recently, researchers have proposed the concepts of national routing, Boomerang Routing, Internet sovereignty, etc. Basically, they are trying to control country level paths of their traffic flows to reduce the risk of being wiretapped. Therefore, it has been necessary and valuable for us to study the geographic properties of Internet routing.

Our study on routing circuitousness shows that the routing circuitousness of our Internet is deteriorating in these years. Routing circuitousness of traffic flows may increase the risk of being wiretapped and circuitous flows may consume more network resources than necessary. Although circuitousness can be a result caused by the design of Internet routing, this continuous trend of deterioration should attract more attentions and its influence and causes should be examined carefully.

Our study on geographical centrality of Internet routing shows the dominance of North America and United States on the data transfer in the Internet. Great British, France and Germany are the other three top countries, but their importances are much less than United States. Our temporal analysis shows the importance of Europe was decreasing comparing with its competitor North America in these years. It also shows almost all countries studied in this paper depend on United States more heavily in 2015 than 2011, with only one exception, i.e. Russia.

We believe this paper is beneficial for people to better understand the routing on the geographical layer map of the Internet.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant no. 61202356 and the National Key Research and Development Program of China under Grant No. 2016YFB0801302.

References

[1] J. Karlin, S. Forrest, J. Rexford, Nation-state routing: censorship, wiretapping, and BGP, CoRR, abs/0903.3218, 2009.
[2] Schengen routing, https://en.wikipedia.org/wiki/Schengen_Routing.
[3] D. D"onni, G.S. Machado, C. Tsiaras, B. Stiller, Schengen Routing: A Compliance Analysis, in: 9th International Conference on Autonomous Infrastructure, Management, and Security (AIMS 2015), in: Lecture Notes in Computer Science, Springer, 2015.
[4] J.A. Obar, A. Clement, Internet surveillance and boomerang routing: a call for Canadian network sovereignty, SSRN Electron. J. (2013), doi:10.2139/ssrn.2311792.
[5] L. Subramanian, V.N. Padmanabhan, R.H. Katz, Geographic properties of internet routing, in: Proceedings of the General Track of the Annual Conference on USENIX Annual Technical Conference, in: ATEC '02, USENIX Association, Berkeley, CA, USA, 2002, pp. 243–259.
[6] P. Du, J.H. Wang, J. Yang, J. Wang, Y. Zhao, Analyzing intercontinental circuitousness to improve the interconnection and routing for isps, in: 2016 International Conference on Information Networking (ICOIN), 2016, pp. 155–160.
[7] P. Mátray, P. Hága, S. Laki, G. Vattay, I. Csabai, On the spatial properties of internet routes, Comput. Netw. 56 (9) (2012) 2237–2248.
[8] B. Eriksson, P. Barford, J. Sommers, R. Nowak, Passive and active measurement: 11th International Conference, PAM 2010, Zurich, Switzerland, April 7–9, 2010. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 171–180.

[9] M.J. Arif, S. Karunasekera, S. Kulkarni, A. Gunatilaka, B. Ristic, Internet host geolocation using maximum likelihood estimation technique, in: Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, 2010, pp. 422–429.
[10] S. Laki, P. Mtray, P. Hga, T. Sebk, I. Csabai, G. Vattay, Spotter: a model based active geolocation service, in: INFOCOM, 2011 Proceedings IEEE, 2011, pp. 3173–3181.
[11] H. Maziku, S. Shetty, K. Han, T. Rogers, Enhancing the classification accuracy of ip geolocation, in: Military Communications Conference, 2012 - MILCOM 2012, 2012, pp. 1–6.
[12] M. Grey, D. Schatz, M. Rossberg, G. Schaefer, Towards distributed geolocation by employing a delay-based optimization scheme, in: Computers and Communication (ISCC), 2014 IEEE Symposium on, 2014, pp. 1–7.
[13] S. Laki, P. Matray, P. Haga, I. Csabai, G. Vattay, A detailed path-latency model for router geolocation, in: Testbeds and Research Infrastructures for the Development of Networks Communities and Workshops, 2009. TridentCom 2009. 5th International Conference on, 2009, pp. 1–6.
[14] B. Huffaker, M. Fomenkov, k. claffy, Drop: Dns-based router positioning, SIGCOMM Comput. Commun. Rev. 44 (3) (2014) 5–13.
[15] V. Giotsas, G. Smaragdakis, B. Huffaker, M. Luckie, k. claffy, Mapping Peering Interconnections to a Facility, in: ACM SIGCOMM Conference on Emerging Networking EXperiments and Technologies (CoNEXT), 2015.
[16] Database: Akamais edgescape, http://www.ip2location.com/.
[17] Digital envoys netacuity, http://www.digitalelement.com/.
[18] Maxminds geoip, http://www.maxmind.com/.
[19] Quova, http://www.quova.com/what/products/.
[20] Hostip free, http://www.hostip.info/dl/index.html.
[21] Ipinfodbfree, http://ipinfodb.com/.
[22] Maxminds geolite city free, http://www.maxmind.com/app/geolitecity.
[23] Software77 free, http://software77.net/geo-ip/.
[24] B. Huffaker, M. Fomenkov, k. claffy, Geocompare: a comparison of public and commercial geolocation databases - Technical Report, Technical Report, Cooperative Association for Internet Data Analysis (CAIDA), 2011.
[25] CAIDA, The caida ucsd ipv4 routed /24 topology dataset 2011–2015, http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.
[26] Maxmind, Geolite2 free downloadable databases, http://dev.maxmind.com/geoip/geoip2/geolite2/.
[27] Internet world stats, http://www.internetworldstats.com/stats.htm.
[28] L. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.

Jessie Hui Wang received her Ph.D degree in Information Engineering from the Chinese University of Hong Kong in 2007. She is currently an assistant professor in Tsinghua University. Her research interests include Internet routing, traffic engineering, network measurement, and Internet economics.

Changqing An received her Master's degree from the Department of Computer Science and Technology at Tsinghua University. She is now an Associate Professor in Tsinghua University. Her research focuses on network measurement and cybersecurity.