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