Technische Universität Berlin
Seminar Internet Measurement
Betreuer: Amir Mehmood

IP Geolocation

Abstract

Geolocation of IP addresses is a nontrivial task which is important for numerous fields of application. The knowledge of the physical location of a user with an assigned IP address is currently being used from credit card fraud protection to online advertising. However, most industrial use approaches to assign ip-addresses or -ranges to a geolocation are currently based on a manually maintained databases which might lead to wrong or outdated information.

Florian Holzhauer
fh-tu@fholzhauer.de
July 9, 2007

1 Introduction

While not being useful for all IP addresses (tunnel-endpoints or mobile nodes, for example), most IP addresses can be traced automatically to their location with an inaccurarcy of several hundreds of kilometres. This might appear high at first, but judged by the fact that it is e.g. sufficient for credit card fraud protection to know the country the user is currently being located, this is a tolerable inaccuracy.

Several examinations developed different mechanisms of automatically geolocation, using a set of servers with known location to triangulate the IP address, using provided location information of the target or topology hints in the router naming scheme. The paper "Towards IP geolocation using Delay and Topology Measurements" covers most of these mechanisms and tries to combine the different measuring methods to maximise the accuracy of the algorithm.

Most of the triangulation methods presented in the paper are easily adaptable to different fields of use, like location of users in a 802.11 environment or a GSM environment. This outline will present an overview over usage, problems, accuracy and implementation variants of geolocation.

Chapter 2 will describe different scenarios where geolocation techniques are currently being used, Chapter 3 discusses semiautomatic approaches to geolocating like manually maintained databases [10], or the usage of reverse dns names[2]. In Chapter 4 features the topic of this paper, automatic approaches to geolocating an IP address - first by a description of already used like GeoPing [11]. It then briefly summarizes the evaluation and comparison of the different algorithms in [6] in Chapter 5, the conclusions are to be found in Chapter 6.

2 Use-Cases

The importance of location accuracy depends very much on the use of geolocation. For some fields of use it is sufficient to have the country the user is being located in, others depend a much higher resolution.

2.1 Online Advertising

It is essential in the advertising business to advertise to the target audience as precise as possible. Online Advertising Providers like Google Adwords try to provide regional targeting [1], which need up to date and precise correlation of the visitors IP address and physical location. Regional Targeting is most useful for services which are available only in a specific region.

2.2 Fraud Protection

Credit Card providers use geolocation to prevent online credit card fraud: The same credit card number used at terminals in different regions or countrys within a small time spam is considered to be an indicator of different persons using the same credit card data, which is not the usual use case for a private credit card.

According to maxmind LLC, a company providing geolocation in combination with other security measures, another use case is to minimize credit card fraud: "If a merchant only wants to sell to consumers in US, Germany, and Japan, they can do so by rejecting orders that were not placed from IP addresses from those countries." [9]

2.3 Legal Issues

Several laws outlaw the use of some online services within a specific region. Popular examples are according to Quova, a provider of geotargeting services: "Pharmacies are not allowed to distribute drugs across national borders. Software and hardware vendors have to comply with OFAC restrictions. And US online gaming firms are not allowed to serve bets to residents of the United States." [13] Similar scenarios apply for copyright matters, so are most TV stations probihited to provide their program outside of their originating country.

2.4 Emergency Calls

Emergency calls are sometimes placed by kids, persons in panic or persons who get unconscious during the phone call: All situations in which it is critical to have accurate and specific information about the callers location. Alerted by several complaints, the FCC is forcing Voice over IP providers to provide 'E911' services, which consist of several information including a callers location [3].

With the gaining importance of voice over IP telephony, this is another usage for geolocation IP addresses with a need for very good approximation of the real location — while approximity of some hundred kilometres might be good enough for most other use cases, this scenario needs very specific numbers.

3 Non-automated Geolocation Techniques

Some of the approaches to geolocating consist of manually maintained work or are at least manually assisted.

3.1 Problems

IP addresses are not static — they can be reassigned, relocated within the same provider or being forwarded using mechanism like vpns or other tunnels. A database consisting of geolocational information about IPs hence has to be updated and maintained on a regular base. Due to the large and increasing number of used IP addresses, this is a task which is nearby impossible to be mantained manually.

3.2 Manually maintained

The most popular approach to geolocation is maintaining a manually created database, like "GeoIP" from MaxMind [10]. A stripped-down version of their commercially available database is given out for free with several tools and apis, and is quite popular within the open source community. Several other free libraries like [8] exist, too.

As already described in section 2.1, this is a rather problematic approach. Paying workers to maintain such a database is very costly. Using community created databases like [8] or the database gained with web2.0 communities like [12] minimizes the costs, but increases the risk of accidentally or intentionally entered wrong data.

3.3 Hints

Several location methods do not qualify as a stand alone method of geolocating an IP because of being not well spread, relying on other information or infrastructure, or are only useful as an addition to the latter described location methods. While this information is not reliable or might be out of date, this still provides valuable hints to verify our estimates.

3.3.1 Provider information

There are several ways where providers sometimes place information regarding the location of a router within the route to the target IP or the IP itself. One of those mechanisms described in [6] are DNS-location-entries [2], usually manually provided by the provider. Another quite common issue is that most providers name the reverse DNS resolution of their routers with some abbreviation of their location, based on airport- or city codes. For example, GiE5-1.ffmxs11.ix.ffm.spxs.net quite obviously is being located in Frankfurt/Main, Germany. However, not all abbreviations are unique, so this is nothing which can be handled without human interaction.

Figure 1: Example of Measurement-Setup — (a) Landmarks, (b) Targets

3.3.2 Heuristics

Using heuristics might gain additional help in providing further verification. As sketched in the paper, mapping the probability of an address location to areas with higher population might be helpful, but can be misleading, too. It is therefore not further investigated in this paper.

4 Automatically generated information

The now described approaches cover automatic discovery of the location of an IP address. While this is obviously more useful for an up-to-date and cheap approach, several other limitations of geolocation still apply. Stations being online via 802.11 Networks, GSM, Satellite based internet or other mobile solutions can move every day or even while being traced. The following mechanisms and algorithms to obtain geolocational information about an IP therefore apply usually only for fixed land lines — although most of the automated approaches like delay measurement can be adopted to geolocate mobile user using the measurement of the delay within underlying transport protocol from the base stations with their known location.

All of the now following measurements rely on a large set of so-called landmarks, hosts which can take several measurements to the target points and have a known geological location. Figure 1 shows the so called university setup used for evaluation in [6].

4.1 Delay-Measurement based

4.1.1 GeoPing

GeoPing assumes that landmarks experiencing similar delays to a target are nearby located to each other. [11] By probing from every landmark, a delay vector is being built and used to calculate which landmark is the closest one to the target. The target is then mapped to the location of that landmark. GeoPing features the usage

of so called passive landmarks which cannot perform measurements to the target. Passive landmarks are servers where the exact location is already known, but can't be used to run own tests, instead, they are used for reference measurements from active landmarks to further eradicate errors caused by routing delays. It is therefore possible to use servers as a probe without deploying an application there.

4.1.2 Shortest Ping

Shortest Ping is a simple approach to delay based measurement, and technically related to GeoPing: Every target is assigned to the closest landmark according to the round trip time.

4.1.3 Constraint-based geolocation

Constraint-based geolocation (CBG) starts with a similar approach like GeoPing [5]. Delay is being measured from all landmarks, and then the location of the target is narrowed down by triangulation: A circle is calculated around every landmark with a radius of the estimated distance. By intersecting these circles, the area the target is assumed to lie in is narrowed down (Figure 2). To relate delay to distance, CBG has to perform initial measurements between all landmarks.

4.1.4 Speed of Internet

Speed of Internet (SOI) is a simplified approach to CBG. One of the disadvantages of CBG is the initial measurement between all known landmarks to create an upper boundary for the delay/distance relation. SOI chose to calculate this upper boundary on a general base. The highest technial possible boundary, which means the farmost distance data can travel within an amount of time, is the speed of data

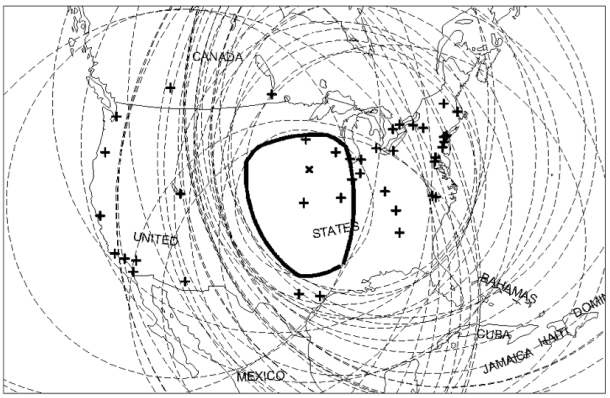

Figure 2: Intersected regions — used by CBG

within fibre cables — which is about $\frac{2}{3}$ of the speed of light in vacuum. However, according to the evaluations in '[6]' this boundary can even be lowered to $\frac{4}{9}$ — less than 0.04% of all measurement are faster in the PlanetLab [7] Setup.

4.2 Topology based Geolocation

Having as much information as possible about the topology of the network we are investigating narrows down the possible physical location of an IP address, hence improving the geolocation process. Topology detection faces several problems. The most important ones for our topic will be described here.

4.2.1 Indirect Routes

An indirect route via one or more routers between the landmark and the target can cause a higher delay between source and target — using delay based geolocation without knowing about the fact we are using an indirect path will result in higher distance estimates than actually present. Figure 5 describes this problem: Assuming a direct path between the landmarks x and y to the target z (a) indicates a bigger distance than first detecting u as an shared router between x and y to z. u therefore has to be located in the intersection of the 'delay radian' from x to u and y to u, and z within the delay vector from u (b). The path from x to z is in this case not as long as first indicated in (a). Detecting a shared route is an indication for such indirect routes. 'A geolocation technique has to therefore take network topology and routes into account in order to capture path-specific latency inflation' [6]. It is important to outline here that for accurate measurements, it has to made sure that the route back to the landmark from the target is the same as the route to it, otherwise the round trip time to the target and the detection to the routers in between might be wrong — which leads to wrong distance calculations.

4.2.2 Hop Locations

As the previous scenario already suggests, calculating not only the location of the target but also the location of the intermediate routers improves the accuracy of the algorithm. If an intermediate router is being located correctly, it even can serve as another additional landmark for all targets behind it. 'In order to achieve a consistent and more accurate solution, a geolocation technique has to simultaneously geolocate the targets as well as routers encountered on paths from landmarks to other landmarks or targets.' [6].

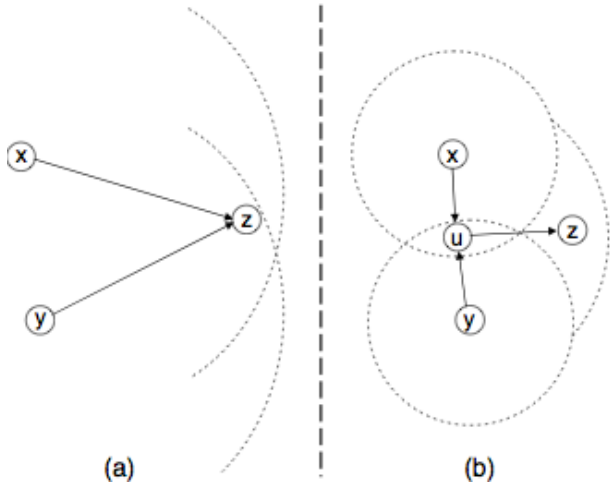

Figure 3: Shared routers, indirect paths.

4.2.3 Clustering

Most routers have more than one network interface. To improve accuray, it is important to find out 'aliased' interfaces — several interfaces representing the same router and the same geological location. Figure 4 describes such a scenario: u and v are aliased interfaces, the accuracy is improved from (a) to (b) where this is detected. 'To tightly fix the feasible locations of routers, a geolocation technique must use measurements to extract existing structural constraints, including by identifying collocated interfaces.' [6] This detection is realized using methods like Mercator [4] and Ally [14].

4.3 Last router

Geolocating the target might not always be useful. In a classic scenario, the user's IP address as the target is rather difficult to geolocate because it is only reachable via one uplink router, it is impossible to geolocate this target via triangulation from different landmarks. [6] thus recommends to geolocate the last router before the final target. Routers are usually multihomed and reachable via different routes, which makes triangulation easier, and is sufficient for most use cases.

5 Evaluation

A comparison of all those algorithms, standalone and combined can be found in length in [6], tested on three different US-sets of landmarks and targets. According

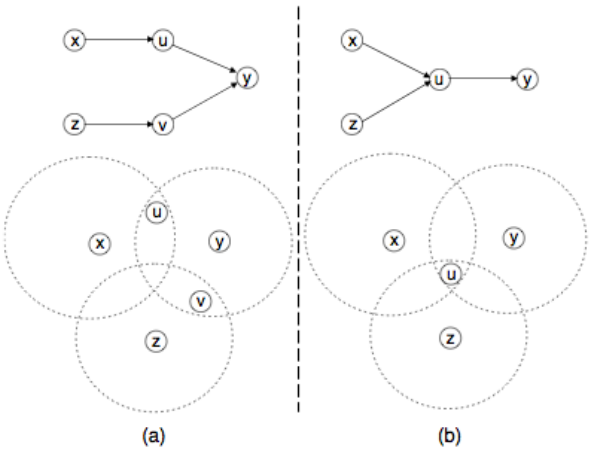

Figure 4: Detection of aliased interfaces.

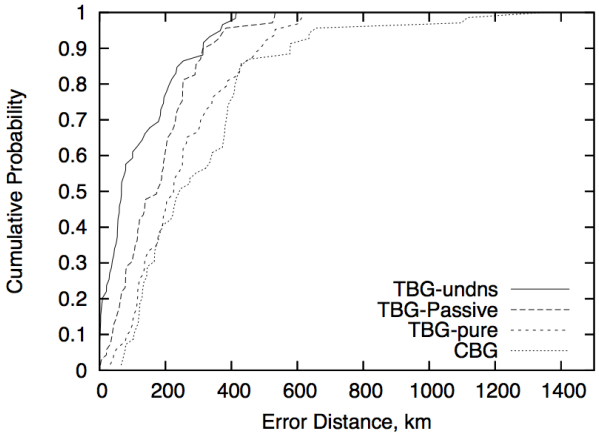

Figure 5: Comparison of CBG and variations of the TBG algorithm.

to the evaluation there, topology based algorithms (tbg) eliminate several disadvantages of delay-only-based algorithms like constraint based geolocation (cgb). While CBG performs very good when the target is close to a landmark, the advantages of TBG become obvious when dealing with targets distinct from all landmarks. Furthermore, according to the authors, CBG 'requires landmarks that completely encircle the target'. TBG can locate targets outside this hull, but has its own disadvantages: It needs sufficient structural constraints on the target. This disadvantage can be minimized using location hints and passive landmarks. Another disadvantage is the need for additional measurements to discover the network topology in advance.

To give an impression of the location accuracy, it might be interesting that the median error in one of the benchmarked datasets in the testings was about 689 km for the SOI-algorithm, 749 km for CBG, and 194 km for 'TBG-pure', which means TBG without the use of passive landmarks and location hints. Other TBG-variations used in the paper are 'TBG-passive', which adds the usage of passive landmarks as reference points and 'TBG-undns', using passive landmarks and DNS validation hints. A comparison of the several TBG-Variations compared to CBG was done on a dataset of 8321 interfaces of several universities (**??**), 2392 alias interface pairs were detected. It is noteworthy that the researchers did not use location hints for .edu tlds, which resulted in DNS-hints for 5509 interfaces. The mean (median) estimation error in this so called 'university dataset' was 138 km (67 km) for TBG-undns, 178 km (176 km) for TBG-passive, 253 km (225 km) for TBG-pure and 296 km (228 km) for CBG.

6 Conclusion

It becomes obvious that the described algorithms are useful for most use cases but the 911-scenario (locating a user placing a emergency call via VoIP), where a median error of several hundreds of kilometers is pretty useless. While TBG provides the highest accuracy, it is noteworthy that it is not always the best choice due to its need to initial measurements. As outlined in the paper, those measurements might take place beforehand, so it does not necessarily impact the speed of geolocation. It depends on the specific use which algorithm might apply best.

References

[1] Google Adwords. Google adwords: Regional and local targeting, 2007.

[2] Christopher Davis. Dns loc: Geo-enabling the domain name system, 2007.

[3] FCC. Fcc consumer advisory: Voip and 911 service, 2007.

[4] Ramesh Govindan and Hongsuda Tangmunarunkit. Heuristics for internet map discovery. In *IEEE INFOCOM 2000*, pages 1371–1380, Tel Aviv, Israel, March 2000. IEEE.

[5] Bamba Gueye, Artur Ziviani, Mark Crovella, and Serge Fdida. Constraint-based geolocation of internet hosts. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 288–293, New York, NY, USA, 2004. ACM Press.

[6] Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards ip geolocation using delay and topology measurements. In *IMC '06: Proceedings of the 6th ACM SIGCOMM on Internet measurement*, pages 71–84, New York, NY, USA, 2006. ACM Press.

[7] Aaron Klingaman, Mark Huang, Steve Muir, and Larry Peterson. PlanetLab Core Specification 4.0. Technical Report PDN–06–032, PlanetLab Consortium, June 2006.

[8] Thomas Mack. open geo coordinates database, 2007.

[9] MaxMind. Maxmind minfraud whitepaper.

[10] MaxMind. Maxmind llc geoip database, 2007.

[11] Venkata N. Padmanabhan and Lakshminarayanan Subramanian. An investigation of geographic mapping techniques for internet hosts. *Proceedings of SIGCOMM'2001*, page 13, 2001.

[12] Plazes. Plazes website, 2007.

[13] Quova. Compliance, 2007.

[14] N. Spring, R. Mahajan, and D. Wetherall. Measuring isp topologies with rocketfuel, 2002.