Bringing location to IP Addresses with IP Geolocation

Jamie Taylor, Joseph Devlin, Kevin Curran

School of Computing and Intelligent Systems
University of Ulster, Magee Campus, Northland Road, Northern Ireland, UK
Email: kj.curran@ulster.ac.uk

Abstract* - IP Geolocation allows us to assign a geographical location to an IP address allowing us to build up a picture of the person behind that IP address. This can have many potential benefits for business and other types of application. The IP address of a device is unique to that device and as such the location can be narrowed down from the continent to the country and even to the street address of the device. This method of tracking can have very broad results and can sometimes only get an accurate result with some input from the user about their location. In some countries laws are in place that state a service can only track you as far as your country without your consent. If the user consents then the service can view your ISP's logs and track you as accurately as possible. The ability to determine the exact location of a person connecting over the Internet can not only lead to innovative location based services but it can also dramatically optimise the shipment of data from end to end. In this paper we will look at applications and methodologies (both traditional and more recent) for IP Geolocation.

I. INTRODUCTION

IP Geolocation is the process of obtaining the geographical location of an individual or party starting out with nothing more than an IP address [1]. The uses (both current and potential) of IP Geolocation are many. Already, the technology is being used in advertising, sales and security. Geolocation is the identification of the real-world geographic location of an Internet-connected computer, mobile device, website visitor or other. IP address Geolocation data can include information such as country, region, city, postal/zip code, latitude, longitude and time zone [1]. Geolocation may refer to the practice of assessing the location, or to the actual assessed location, or to locational data. Geolocation is increasingly being implemented to ensure web users around the world are successfully being navigated to content that has been localised for them. Due to the '.com' dilemma, most companies are finding that more than half of the visitors to their global (.com) home pages are based outside of their home markets. The majority of these users do not find the country site that has been developed for them. Companies such as Amazon have introduced geolocation as a method of dealing with this problem [2].

There are organisations that are responsible for allocating IP address. The Internet Assigned Number Authority(IANA) is responsible for allocating large blocks of IP addresses to the following five Regional Internet Registries(RIR) that serve specific regions in the world: AfriNIC (Africa), APNIC (Asia/Pacific), ARIN (North America), LACNIC (Latin America) and RIPE NCC (Europe, the Middle East and Central Asia). These RIR's then allocate blocks of IP addresses to Internet Service Providers (ISP) who then allocates IP addresses to businesses, organizations and individual consumers. Using the above information IP addresses can be broken down into graphical locations within few steps but to get a more accurate result than that the user may have to provide additional details to aid the process. In some cases, this practice will become more efficient the more it is used, a user's location can be tracked by closely matching their IP address with a neighbouring IP address that has already been located. Many businesses have been started up just by hosting large databases of IP address to allow services to apply this technology with varying degrees of efficiency and accuracy. There are many methods of tracking a device such as GPS and cell phone triangulation but Geolocation, the least accurate, is becoming popular among website owners and government bodies alike

The foundation for geolocation is the Internet protocol (IP) address, a numeric string assigned to every device attached to the Internet. When you surf the web, your computer sends out this IP address to every website you visit. IP addresses are not like mailing addresses. That is, most are not fixed to a specific geographic location. And knowing that a particular ISP (Internet Service Provider) is based in a particular city is no guarantee that you'll know where its customers are located [3]. That is where geolocation service providers come in. Geolocation service providers build massive databases that link each IP address to a specific location. Some geolocation databases are available for sale, and some can also be searched for free online**.** As the IP system is in a constant state of flux, many providers update their databases on a daily or weekly basis. Some geolocation vendors report a 510% change in IP addresses locations each week. Geolocation can provide much more than a geographic location. Many geolocation providers supply up to 30

doi:10.4304/jetwi.4.3.273-277

data fields for each IP address that can help to further determine if users really are where they say they are. These may include country, region, state, city, ZIP code, area code; Latitude/longitude; Time zone; Network connection type and domain name and type (i.e. .com or .edu). Not every IP address accurately represents the location of the web user. For example, some multinational companies route Internet traffic from their many international offices through a few IP addresses, which may create the impression that some Internet users are in, say, the UK when they are actually based in France. If someone is using a dial-up connection from Ireland back to their ISP provider in the France, it will appear like they are in the France. There are also proxy services that allow web users to cloak their identities online, a few geolocation providers however have introduced technology that can look past these proxy servers to access the user's true location. In addition, some providers can now locate, down to a city-street level, people connecting to the Internet via mobile phones or public Wi-Fi networks. This is accomplished through cell tower and Wi-Fi access point triangulation [4].

Here, we will be looking at this technology in more detail and what it could mean for us and our lives going forward. This paper is structured as follows. In Section 2, we look in more depth at the applications for IP Geolocation (both current and potential). Section 3 then presents a number of IP Geolocation methods, starting with more 'traditional' methods before progressing to those more recent and 'hybrid' in nature. In section 4, we outline some methods for avoiding IP geolocation and we conclude our discussion in Section 5.

II. IP GEOLOCATION USAGE

Localization is the process of adapting a product or service to target a specific group of users. These changes can include the look and feel of the product, the language and even fundamental changes in how the service or product works. Many global organisations would like to be able to tailor the experience of a website to the types of users viewing it as it can have a significant impact on whether or not a user will use your service. The ability to gather useful metrics increases when you add in the fact you can tell where your customers are from. Take for example Google. Google provides localized versions of its search engine to almost every country in the world. Using Geolocation, they can select the correct language for each user and alter their search results to reflect more accurately what it is the user is actually searching for. If Google so chose they could even start to omit certain results to comply with national laws. Google ads use this feature heavily by making sure that local businesses can reach people in their area so as to increase the impact of the advertising. This localization of websites is becoming increasingly popular and Geolocation is a tool that grants the ability to easily find out which version of your website to show.

Other websites use localisation in the opposite way. Instead of attempting to increase the use of the site by accommodating worldwide users some websites would use a user's location to ensure that they cannot access the website or its content. This practice is most common on sites that host copyrighted content such as movies, TV shows or music. An example of this is the BBC iPlayer. This service cannot be accessed in the USA for example, as the BBC iPlayer will not allow anyone with an IP address outside the UK to view the content. Online gaming/gambling websites use Geolocation tools to ensure that they are not committing crimes in countries where gambling is illegal [5].. An example of this is www.WilliamHill.com. This website filters out American users to avoid breaking laws in that country. In Italy, a country where gambling is illegal, you will only be granted a license to host a gambling website if you apply Geolocation tools to restrict access to the site by Italians. MegaUpload.com in 2012 was involved in a legal dispute with regard to their facilitating of copyright infringement. To try to avoid charges such as those the company, who have all their assets in Hong Kong, made sure to use Geolocation tools to filter out anyone in Hong Kong from using their services. This meant that MegaUpload.com was committing copyright infringement in every country in the world except Hong Kong.

IP Geolocation has a vast array of both current and potential uses and areas of application. Of course, the accuracy (or granularity) needed varies from application to application. Through the use of IP Geolocation, advertisements can be specifically tailored to an individual based on their geographical location. For example, a user in London will see adverts relative to the London area, a user in New York will see adverts relative to the New York area and so forth. Additional information such as local currency, pricing and tax can also be presented. A real life example of this would be Google AdSense. As one may imagine, the accuracy needed for this is considerable; we would need a town or (even better) a street as opposed to say a country or state in order to provide accurate information to the user.

As the online space continues to become *the place* to do business, issues once thought to be solved now rear their heads again. For example, DVD drives were region locked to prevent media being played outside the intended region, but problems exist in combating this resurging issue in the online space. Other examples of content restriction include the enforcement of blackout restrictions for broadcasting, blocking illegal downloads and the filtering of material based on culture. Content localisation on the other hand, is working to ensure that only relevant information is displayed to the user. The accuracy needed for an application shows a visitor from Miami, Florida dressed in beach attire instead of parkas is less than that needed for advertising. Here, geolocation at the country level is normally sufficient to ensure that users from one country cannot access content exclusive to another country for example.

Businesses can often struggle to adhere to national and regional laws due to the degree of variance. Failure to comply with these laws however can result in financial penalties or even prison time. Advertising for instance, can be subject to tight control such as what can be advertised where, when and if the product or service in question can be advertised in a particular location at all. Indeed, even the above examples of content restriction are often done to comply with legal requirements. In addition, there is the need to avoid trading with countries, groups and individuals black-listed by government. Quova[1] provide us with the OFAC (Office for Foreign Assets Control – United States) and the need to comply with its economic and trade sanctions as an example of this. IP Geolocation offers us a powerful tool to help us comply with these legal requirements. However, to use IP Geolocation effectively in this scenario, we would need a state-level accuracy as laws can vary from state to state.

IP Geolocation can also offer much to those in security. IP Geolocation is used as a security measure by financial institutions to help protect against fraud by checking the geographical location of the user and comparing it with common trends. In the field of sales, user location can be compared with billing address for example. MaxMind[2] is one such group offering products such as MinFraud which provides relevant information about the IP's historic behaviour, legitimate and suspicious and attempts to detect potential fraud by analysing the differences between the user location and billing address.

III. METHODS OF IP GEOLOCATION

A common approach to IP Geolocation is to create and manually maintain a database containing relevant data. These Non-automated methods, (i.e. those relying on some form of human interaction or contribution) can be undesirable. Problems include the fact that IP addresses are dynamically assigned and not static and therefore the database requires frequent updating (potentially at considerable financial cost and the risk of human error). The switch from IPv4 (2^32 possible addresses) to IPv6 (2^128 possible addresses) increases the challenge exponentially. One approach is to rely on delay measurements in order to geolocate a target. It should be noted however that these approaches rely on a set of 'landmarks', where a landmark is some point whose location is already known. A common way often used to construct this set of landmarks is to take a subset of nodes from the PlanetLab[3] network (consisting of more than 1000 nodes).

Delay Based Methods

Constraint Based Geolocation (CBG) is a delay-based method employing multilateration (estimating a position using some fixed points) [6]. The ability of CBG to create and maintain a dynamic relationship between IP address and geographical location is one of the methods key contributions to the IP Geolocation process, since most preceding work relied on a static IP address to geographical location relationship. To calculate this distance, each landmark measures its distance from all other landmarks. A bestline is then created where a bestline is the least distorted relationship between geographic distance and network delay.

A circle then emanates from each landmark, the radius of which represents the targets estimated distance (calculated above) from that landmark. The area of intersection is the region in which the target is believed to reside; CBG will commonly guess that the target is at the centroid of this region. The area of this region is an indication of confidence, the smaller the area, the more confident CBG is in its answer, and a larger area implies a lower level of confidence.

Speed of Internet (SOI) [7] can be viewed as a simplification of CBG. Whereas CBG calculates a distance-to-delay conversion value for each landmark, SOI instead uses a general conversion value across all landmarks. This value is 4/9c (where c is the speed of light in a vacuum). Numerous delays (such as circuitous paths and packetization) prevent data from travelling through fibre optic cables at its highest potential speed (2/3c). Therefore it is reasoned that 4/9c can be used to safely narrow the region of intersection without sacrificing location accuracy. Shortest Ping is the simplest delay-based technique. In this approach a target is simply mapped to the closest landmark based on round-trip time (aka ping time). Delay-based methods rely on the distance between the target and its nearest landmark. This is a good predicator of the estimation error. Round-trip time is also a good indication of the error; delay-based methods work well when the RTT is small and performance deteriorates relative to the increase in RTT. Having to effectively take the network as is, feeling your way around with delay measurements as opposed to being able to map it out to potentially improve accuracy is something we'd be keen to overcome. As we will see, using topology information and other forms of external information can greatly increase accuracy.

Topology-based Geolocation (TBG)

The methods here attempt to go beyond using delay measurements as their sole metric. Some seek to combine traditional delay measurements with additional information such as knowledge of network topology and other additional information; some even attempt to recast

[1] www.quova.com
[2] www.maxmind.com
[3] www.planet-lab.org/

the problem entirely. The reliance of delay-based methods upon a carefully chosen set of landmarks is a problem [7]. Topology-based Geolocation (TBG) however uses topology in addition to delay-based measurements to increase consistency and accuracy. This topology is the combination of the set of measurements between landmarks, the set of measurements between landmarks and the target (both measurements obtained by traceroute) and the structural observations about collocated interfaces. The target is then located using this topology in conjunction with end to end delays and per hop latency estimates. When presented with a number of potential locations for a target, TBG will map the target to the location of the last constrained router.

It should be noted however that TBG incurs some overheads that simple delay-based methods do not. TBG must first construct its topology information and an additional overhead can be found in refreshing this information to ensure it is up to date and accurate. However the authors point out that this topology information can be used for multiple targets and that this overhead need not necessarily apply to every measurement one may wish to make. There are three main variants of TBG. These are

1) TBG-pure, using active landmarks only
2) TBG-passive, using active and passive landmarks
3) TBG-undns, using active and passive landmarks in conjunction with verified hints

Once successfully located, intermediate routers can be used as additional landmarks to help locate other network entities.

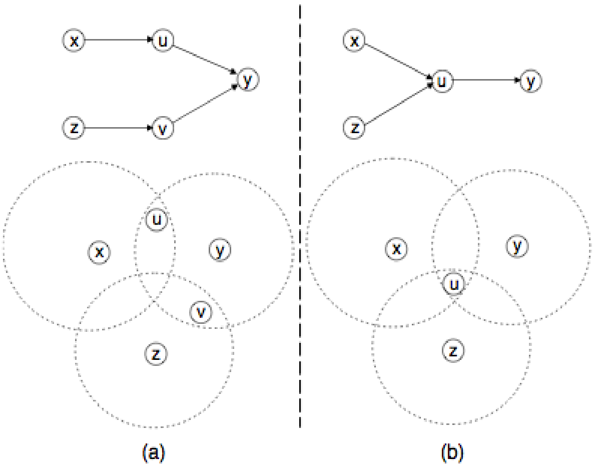

Figure 1: Identifying & clustering multiple network interfaces [7]

In order to accurately determine the locations of these intermediate routers with confidence and be able to use them effectively, we must record position estimates for all routers encountered so we can base our final position estimate on as much information as possible. For instance, in trying to geolocate a router that is one hop from a given point and multiple hops from another given point, we need to record all routers we encounter which will allow us to determine its position with more accuracy. In other words, a geolocation technique has to simultaneously geolocate the targets as well as routers encountered [7]. Discovering that a router has multiple network interfaces is a common occurrence. Normally these interfaces are then grouped (or clustered) together (this process of identification and resolution is also known as IP aliasing), otherwise we falsely inflate and complicate our topology information. Part *a* of Figure 1 shows two routers *u* and *v* which are in fact multiple interfaces for the same physical router. In part *b* we see how the topology has been simplified by identifying *u* and *v* and clustering them.

Web Parsing Approach

Another approach to improve upon delay-based methods is through the use of additional external information beyond inherently inaccurate delay-based measurements [8]. This can be achieved through parsing additional information from the web. Prime candidates therefore are those who have their geographic location on their website. The method seeks to extract, verify and utilize this information to improve accuracy. The overall system is made up of two main components, first is a three-tier measurement methodology, which seeks to obtain a targets location. The second part is a methodology for extracting and verifying information from the web then used to create web-based landmarks.

This three-tier measurement methodology uses a slightly modified version of CBG (where 4/9c is used as an upper bound rather than 2/3c) to obtain a rough starting point. Tiers 2 and 3 then bring in information from the web obtained by the second component to increase the accuracy of the final result. The information extraction and verification methodology relies on websites having a geographical address (primarily a ZIP code). This ZIP code combined with a keyword such as university or business is passed to a public mapping. If this produces multiple IPs within the domain name then they are to be grouped together and refined during the verification process. Assuming one has used a public search tool as suggested: the first stage of verification is to remove results from the search if their ZIP code does not match that in the original query. In cases where the use of a Shared Hosting technique or a CDN (Content Delivery Network) results in an IP address being used for multiple domain names the landmark (and subsequently the IP) is to be discarded. Finally, in the case where a branch office assumes the IP of its headquarters: compare ZIP codes again to confirm its identity as a branch and subsequently remove it. As with TBG described above, the method presented here also succumbs to certain overheads that delay-based methods are able to avoid. The measurement stage has a delay of 1-2 seconds for each measurement made. This is the result of 8 RTTs (Round Trip Time); 2 of which are performed in the first tier and 3 in both the second and third tiers. The verification stage incurs an overhead for each ZIP code considered as all landmarks for each ZIP are cached. However, this will only require occasional updates and thus does not affect each and every search. [9] attempt to improve the accuracy of IP Geolocation by broadening the scope of information considered through casting IP Geolocation as a *machine learning-based* classification problem. Here a Naive Bayes classifier is used along with a set of latency, hop count and population density measurements. Each of these metrics/variables can be assigned a weight to affect how it will influence and inform the classifier. Results are classed in quintiles, with each quintile representing 20% of the target IPs and a level of confidence in the results within that quintile.

IV. GEOLOCATION EVASION (CYBERTRAVEL)

With Geolocation restrictions becoming more popular internet users are finding ways to evade these restrictions. Every country has its own laws that they are applying to cases involving Geolocation, but those laws were not written with the technology in mind. Cybertravel in a phrase, admittedly almost unknown but apt, that refers to evading Geolocation, GPS and other similar tracking technologies by pretending that you are in a real world location that you are not. Cybertravel is not the same as making yourself anonymous as the latter is about making your location unknown and the former is about providing an incorrect location. One way to do this is to alter your IP address to make it seem that you are from another region. Many people use this evasion technique to access content that is restricted. It is popular with people who are trying to access websites hosting copyrighted TV shows. Another less popular way to cybertravel is to gain remote access another device that is physically in the region that you want to access the internet from. This way you are not actually altering you IP; it is as if you had physically travelled to that region and accessed the internet from there. Services like TOR[4] provide you with internet anonymity. Actively trying to hide your identity or location can have the result that websites cannot determine your region and thus may not allow you access to their content at all which is why cybertravel is the method of choice for accessing region restricted content.

Services exist that allow a user to pay a monthly fee in return for an IP from a particular region. An example of this is www.myexpatnetwork.co.uk. This company allows users from outside the UK to gain a UK IP address. This company only deals in the GBP currency and is marketed to UK residents. The company is not breaking any laws by 'leasing' these IP addresses and because they are marketing at UK residents who are abroad they may believe they are covered from advertising a Geolocation evasion tool. This service description may not stand up in court as a large portion of their customers are likely non UK residents looking to access UK only content. Evasion of Geolocation has not become a major issue at the moment. As with many issues like this, most organisations do not care until services like myexpatnetwork become popular and so easy to use that a serious financial loss looms. Governments are starting to pay attention to this issue now as they start to understand the difficulty of enforcing their laws against companies and people outside their jurisdiction who commit crimes on the internet.

V. CONCLUSION

We have provided an overview of IP Geolocation applications and methodologies both traditional and those that attempt to push the envelope. The methodologies presented here vary both in their complexity and accuracy; as such, we cannot claim any one method as the ideal solution. The optimal approach is therefore highly sensitive to the type of application being developed.

REFERENCES

[1] Lassabe, F. (2009). *Geolocalisation et prediction dans les reseaux Wi-Fi en interieur*. PhD thesis, Université de Franche-Comté. Besançon

[2] Brewster, S., Dunlop, M., 2002. Mobile Computer Interaction. ISBN: 978-3-540-23086-1. Springer.

[3] Furey, E., Curran, K., Lunney, T., Woods, D. and Santos, J. (2008) *Location Awareness Trials at the University of Ulster,* Networkshop 2008 - The JANET UK International Workshop on Networking 2008, The University of Strathclyde, 8th-10th April 2008

[4] Furey, E., Curran, K. and McKevitt, P. (2010) *Predictive Indoor Tracking by the Probabilistic Modelling of Human Movement Habits*. IERIC 2010- Intel European Research and Innovation Conference 2010, Intel Ireland Campus, Leixlip, Co Kildare, 12-14th October 2010

[5] Sawyer, S. (2011) EU Online Gambling and IP Geolocation, Neustar IP Intelligence, http://www.quova.com/blog-2/4994/

[6] Gueye, B., Ziviani, A., Crovella, M. and Fdida, S. (2004) Constaint Based Geolocation of internet hosts. In IMC '04. Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, pp: 288 – 293.

[7] Katz-Bassett, E., John, J., Krishnamurthy, A., Wetherall, D., Anderson, T. and Chawathe, Y. (2006) Towards IP Geolocation Using Delay and Topology Measurements. In ICM '06. Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, pages 71 – 84.

[8] Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A. and Huang, C. (2011) Towards Street-Level Client Independent IP Geolocation. In NSDI'11. Proceedings of the 8th USENIX conference on networked systems design and implementation, pp: 27-36

[9] Eriksson, B., Barford, P., Sommers, J. and Nowak, R. (2010) A Learning-based Approach for IP Geolocation. IN PAM'10 Proceedings of the 11th international conference on Passive and active measurement, pp: 171 – 180.

[4] https://www.torproject.org/