Geolocation by IP Address: A Summary

June 2018

Andrew Young
R&D Data Scientist, Manager
Neustar Inc.
andrew.wong@neustar.biz

***Abstract*—A single IP address may serve many devices and thus demonstrates a "one-to-many" relationship. Examples of such high traffic IP addresses include coffee shops, universities, library, hospitals, airports and the list goes on. Similarly, a single device may be linked to many IP addresses as its owner goes about his or her day. Consequently, IP addresses and mobile devices exhibit a "many-to-many" relationship. To complicate matters, your home may have many IP addresses associated with it over some set period of time. Thus, we have a many-to-many-to-many relationship. Resolving this many device to many IP address to many home address relationship into a many device, one IP address, one home address (many-to-one-to-one) relationship can be complicated. But despite the challenges involved, solving the aforementioned problem poses a lucrative reward. It improves Neustar's high-profile OneID project, which seeks to build a database of federated identities for CRM, fraud detection, market advertising segmentation and other purposes. OneID provides the infrastructure for most of Neustar's products/services, most of its existing revenue stream and is considered to be the primary revenue growth engine for the future. In the paper below, the nuances of IP addresses and the unique solutions they enable are explored.**

I. Introduction

Using IP addresses to link user records helps build a unified data set with more attributes for a given entity. More information helps with better segmentation and provides a more holistic customer profile.

II. Background & Motivation

IP address geolocation is used for online fraud prevention, localized content, privacy compliance (i.e. GDPR), targeted advertising & more. With respect to online advertising, better online identity resolution allows display of more relevant offers to specific customer segments, better price discrimination and attractive visuals that increase time on a website and drive higher conversion rates. Further segmentation of traffic and insight into customer behavior. Ads can also be targeted based on location, industry and business to improve responses through customized offers and creative. Wasted impressions are reduced and reflected in measurable results. It is a lucrative problem to solve, as evidenced by Neustar's purchase of a company called Quova in late 2010.[13] With the acquisition, Neustar can better enable businesses to geographically locate their website users for advertising and product customization activities.

"Every machine on the Internet has a unique identifying number called an IP Address. The IP stands for Internet Protocol, which is the language that computers use to communicate over the Internet."[15] A protocol is the pre-defined way that someone (e.g. a person or a Web browser) talks with a service they want to use.

As it stands, there are two standards for internet protocol (IP) addresses: IPv4 (v4 means version 4) and IPv6. IPv4 addresses are 32-bit numbers, theoretically providing 232 = 4,294,967,296 IP addresses. For various reasons, there are realistically only about 3.7 billion public IPv4 addresses.[8][12] Due to the rapid growth of handheld device ownership and increasing Internet utilization in countries with historically low Internet participation rates, a pressing concern is the lack of sufficient IP addresses to satisfy demand, a problem also known as IPv4 address exhaustion. Hence, the development of the IPv6 standard by the international open standards group known as the Internet Engineering Task Force (IETF).[10]

IPv6 addresses are 128-bit numbers, theoretically providing 3.4 * 1038 addresses. The IPv6 standard was formalized by 1998 and IPv6 addresses have been available for allocation since 1999, but adoption has been slow due to implementation costs (e.g. router replacement). It is helpful to understand how a network works, so an example of a network and how two networks communicate is provided below.

> ... Imagine that Company A is a large ISP. In each major city, Company A has a POP [Point of Presence]. The POP in each city is a rack full of modems that the ISP's customers dial into. Company A leases fiber optic lines from the phone company to connect the POPs together. ... Imagine that Company B is a corporate ISP. Company B builds large buildings in major cities and corporations locate their Internet server machines in these buildings. Company B is such a large company that it runs its own fiber optic lines between its buildings so that they are all interconnected. In this arrangement, all of Company A's customers can talk to each other, and all of Company B's customers can talk to each other, but there is no way for Company A's customers and Company B's customers to intercommunicate.
> Therefore, Company A and Company B both agree to connect to NAPs (Network Access Points) in various cities, and traffic between the two companies

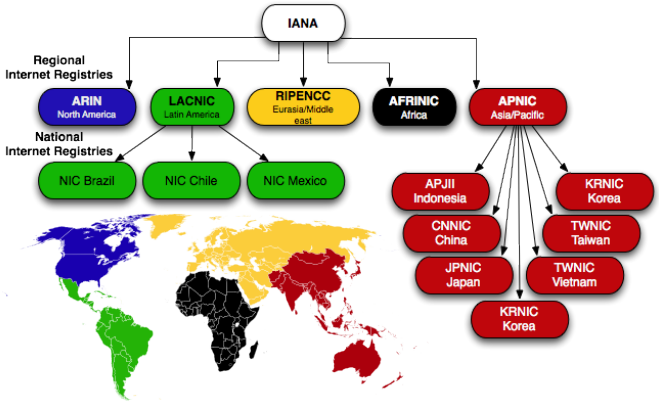


Fig. 1. "Geographic breakdown of the Regional Internet Registries (RIR) and National Internet Registries (NIR) (Retrieved 2016 from the Center for Applied Internet Data Analysis.[15]).

> flows between the networks at the NAPs. In the real Internet, dozens of large Internet providers interconnect at NAPs in various cities, and trillions of bytes of data flow between the individual networks at these points. The Internet is a collection of huge corporate networks that agree to all intercommunicate with each other at the NAPs. In this way, every computer on the Internet connects to every other (Tyson).[15]

III. Specific Questions

Below are some specific questions that will help guide the feasibility analysis of using IP addresses as a source of geographic information, specifically latitude/longitude information.

How does IP address assignment work?

The broad strokes of high-level IP address assignment are well-documented. The department of a US-based nonprofit known as Internet Assigned Numbers Authority (IANA) manages global IP address space allocation and delegates five regional Internet registries (RIRs) to further allocate IP address blocks to local Internet registries (LIRs) (e.g. Internet service providers (ISPs)) and other entities.[18] Figure 1 provides a high-level illustration of the IP assignment process to the world's regions.

It is less clear how public IP addresses are assigned at the household level. ISPs may have different assignment protocol for IPv4 addresses and it is possible there are differences in assignment protocol between IPv4 and IPv6 addresses even for the same ISP. However, information on these assignment mechanisms is not freely available since an exhaustive search revealed nothing.

Does assignment protocol differ between IPv4 and IPv6 addresses?

Uncertain. It appears that the majority of the world still uses IPv4 (Internet Protocol version 4) despite the world approaching the late stages of address space exhaustion. What is clear is that companies that provide IP geolocation services distinguish between the two, offering geolocation information on IPv6 addresses for an additional fee.

How often do public dynamic IP addresses change?

IP addresses are "leased" to residential users. These lease terms have a default length that ranges anywhere from 24 hours to 7 days, at which point they are automatically renewed in most cases. Consequently, some users can have the same public IP address for years and effectively have a static IP address.[5]

Static IPs offer several advantages over dynamic IPs for a subset of the population with specific use-cases. For example, it is beneficial for a small business to have a static IP address so that customers, suppliers and others can reliably connect to the business's server. Additionally, it allows remote access as well as more stability that makes interaction with email or chat servers, database servers, network equipment and VPN providers more convenient.[14]

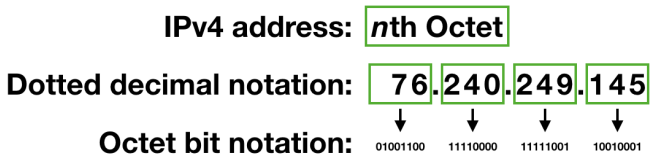
Historically, most people used the Internet for only a few minutes to a few hours per week. Since Internet Service Providers (ISPs) pay for their block of IP addresses, assigning a dynamic IP system allowed them to "lease" the same IP address to multiple customers, decreasing their cost-percustomer while maximizing utility of their service.

As broadband connections have become more common the distinction between dynamic and static IPs has become less well delineated and the majority of connections are "always-on" even when nobody is actively using the Internet.

> When modern ISPs enforce dynamic IPs these days, it may be in part to distinguish between "consumer" and "professional" servicesby reserving static IPs for customers who pay more, it gives customers who need that feature an incentive to upgrade their service level. It can also serve as a deterrent for people abusing their consumer-grade service. Many ISPs, for instance, explicitly prohibit running "servers" on a home Internet connection. If every home user had a static IP, theyd be more inclined to abuse such terms of service. It's also less of a management problem to assign customers dynamic IPs. If you move across town (but within the same ISPs service area), theres no need to re-assign how your static IP is routed; youll just get a dynamic IP that exists in the new neighborhood.[16]

What type of information be extracted from the IP address itself?

While we are interested in geographic information, such is not explicitly tied to an IP address. However, there are tables that are manually maintained on a voluntary basis tying users to specific geographies. For example, there are websites



Address Class	First Octet (decimal)
Class A	1 - 126
Class B	128 - 191
Class C	192 - 223

Fig. 2. Any IPv4 address can be translated into 4 sets of "octets", a unit of information that consists of eight bits. An octet is a binary representation of the more human-friendly IPv4 addresses that we see. The green boxes relate the 1st through 4th octets of an IPv4 address to their corresponding binary representations.

IP Address Class	First Octet (decimal)	Max hosts	Example
Class A	1 - 126	17×10^6	76.240.249.145
Class B	128 - 191	65×10^3	130.240.249.145
Class C	192 - 223	254×10^0	195.240.249.145

Fig. 3. An IPv4 address can also be considered to be composed of two parts: a network ID and a host ID, as distinguished by the blue and yellow boxes. The class of an IP address determines the number of octets comprising its network ID and by extension, its host ID.[1] Universities are usually believed to have Class A IP addresses to allow for the large number of unique IP addresses needed for its student, faculty and visitor population. Residential public IP addresses tend to be class C IP addresses since they rarely need more than 254 unique IP addresses.

where users can look up their IP address and see their predicted geography information with the option to adjust the assessment.

IP addresses consist of two parts: a network ID and a host ID. A network ID may provide clues to the nature of the IP address's owner, though this is not always the case. Some sources claim that universities tend to have class A network IPv4 addresses. However, Oxford University uses multiple class B and class C networks while Stanford has moved to the IPv6 standard and returned its IPv4 block allocation to help delay the complete address exhaustion date.[11] Additionally, the adoption of a new system called Classless Inter-Domain Routing (CIDR) complicates inferences based on this feature.

How does IP work on mobile cellular networks? Does IP change more frequently?

Yes, the IP addresses assigned on a mobile data network change more frequently. According to Charlie Stobert, a self-described professional with 10 years of experience at Cisco and Juniper via Quora:

> The UE (User Equipment) or "mobile phone" is assigned an address by a process similar to Dynamic Host Configuration Protocol (DHCP) by the mobile operator in question. Every time the UE is restarted, or even when moving across an invisible network boundary that you're not even aware of, your IP address is likely to change. Also most operators would have to dynamically use Network Address Translation (NAT) to remap a UE's private address to a public address in order to conserve public IP addresses.

Neustar's Rupert Young, a former Sr. Director of Product Management at Quova (an IP intelligence firm acquired by Neustar), had a similar answer:

> IP data on mobile networks: typically most mobile networks are set-up as a proxy with a small number of public IPs assigned to millions of devices often over multi-state regions, (i.e. AT&T. Verizon has more localized gateways). IPs are usually reassigned to another user within minutes, like after you stop browsing the internet on your mobile browser. So, purely mobile IPs are useless. However, the majority of mobile device usage has been on WiFi networks, where you can trust the IP location to within miles.

Does an IP address contain information that may indicate whether it is a residence or a business?

Yes and no. Businesses are more likely to have a static IP address, so on average, an entity with an IP address that changes less frequently is more likely to be a business. However, as previously mentioned, many dynamic IP addresses are effectively static due to frequent IP address lease renewals in an age of "always-on" connections. Additionally, IP addresses themselves contain "no" information.[2517] To make inference on whether an IP belongs to a residence or a business requires more sophisticated data mining methods. According to Rupert, "Our data does have a home/non-home field that is a decent starting place, however, you also have to layer on usage filters (# of cookies or DID's per IP, for example) to remove IP addresses." High-traffic IP addresses are associated with too many cookies or devices, which may indicate a non-residential IP address.

In terms of lat/long information, is it the location of the customer?

Lat/long information from IP address is highly inaccurate.

It is tied to the predicted city of the customer. Ten different IP addresses tied to the city of Los Angeles may return the exact same lat/long if one uses the same IP geolocation service. Additionally, a single IP address can be matched with varying lat/long addresses depending on the IP geolocation service. Cities can encompass large areas so I was curious how the lat/long of a city was determined. I was unable to find this information, but I'm guessing it may indicate the relative center (i.e. centroid) of the city. Note that there are varying degrees of precision (# of significant figures) associated with lat/long measures and answers may vary based on the geolocation service used.

According to Rupert, "Lat/lon, if you are referring to the data from an IPI look up is the centroid of the zipcode we have mapped the IP to. If you are referring to the lat/lon that comes with a 3rd party API/batch call, sometimes it is the real lat/lon of the device, sometimes it is stuffed data that someone looked up somewhere in the ecosystem from an IP data vendor and are pretending it is the real lat/lon. For the real lat/long, sometimes it is WiFi triangulation (relatively accurate except in-doors or more rural areas) and sometimes it is GPS-based."

IV. Solutions

There are many services that provide geolocation services based on IP addresses with varying degrees of accuracy. Following Neustar's acquisition of Quova, Neustar also has ability to geolocate based on IP addresses. The following tables compare four services in terms of features offered and user-reported accuracy for a 90 day period.

Available Database Fields

	Neustar	MaxMind	IP2Location	IPligence
Continent	X	X	X	X
Country	X	X	X	X
State/Region	X	X	X	X
City	X	X	X	X
Zip Code	X	X	X	
Latitude	X	X	X	X
Longitude	X	X	X	X
ISP	X	X	X	X
Organization	X	X		
Organization Type	X	X		
ASN	X	X		
Net Speed	X	X	X	
Net Type	X	X		
Domain	X	X	X	
Area Code	X	X	X	
Metro Code	X	X	X	
DMA Code	X	X		
Weather Station			X	
Time Zone	X	X	X	X
Proxy	X	X	X	
Subscription	Query	Monthly	Annual	Annual

Fig. 4. "Geolocation services offer a varying number of fields, or features. Retrieved in 2018 from WhatsMyIPAddress[3])."

Geolocation Accuracy

Summary of user reported accuracy of geolocation services over the past 90 days.

Distance	W3C	Neustar	MaxMind	IP2Location
0-2 km	53%	52%	33%	31%
2-10 km	9%	11%	9%	8%
10-25 km	4%	4%	6%	6%
25-50 km	4%	4%	6%	6%
50-100 km	3%	3%	8%	8%
100-250 km	3%	3%	7%	7%
250-500 km	3%	3%	6%	6%
500-1000 km	3%	4%	8%	8%
1000+ km	19%	14%	18%	20%

Fig. 5. "A comparison of user-reported accuracy for four services. Retrieved in 2018 from WhatsMyIPAddress[3])."

How does geolocating by IP address work?

"The primary source for IP address data is the regional Internet registries which allocate and distribute IP addresses amongst organizations located in their respective service regions:

1) African Network Information Centre (AfriNIC)
2) American Registry for Internet Numbers (ARIN)
3) Asia-Pacific Network Information Centre (AP-NIC)
4) Latin American and Caribbean Internet Address Registry (LACNIC)
5) RIPE Network Coordination Centre (RIPE NCC)

Secondary sources include:

1) Data mining or user-submitted geographic location data. For example, a weather web site might ask visitors for a city name to find their local forecast. Another example would be to pair a user's IP address with the address information in his/her account profile.
2) Data contributed by internet service providers.
3) Merging databases from different suppliers.
4) Guesstimates from adjacent Class C range and/or gleaned from network hops.

Accuracy is improved by[4]:[7]

1) Data scrubbing to filter out or identify anoma- lies.
2) Statistical analysis of user submitted data.
3) Utilizing third-party tests conducted by reputable organi- zations

However, many claim geolocation has questionable accuracy. According to a self-described senior-level employee of an ISP on StackOverflow:[6]

> Large IP ranges are allocated as needed by IANA to each of the Regional Internet Registries.[9] The regions are generally continental in size - IP addresses are not assigned on a per-country basis. The RIRs in turn then allocate IP addresses to ISPs, who in turn assign them to end-users. Each of the RIRs maintain a whois server which can be queried to find out

> not only which ISP has been assigned any netblock, but to a certain extent which end-user, and that end-user's address. Note that many ISPs do not fill out this information for every single customer. Hence if you're a residential subscriber of a DSL service, it's likely that the Geo records will give the address of your ISP, and not your own address. The various GeoLocation providers mostly work by mining these whois records. Note that the legality of doing so is something of a gray area ...

V. Future Work

More time can be invested in continuing literature review and research. Additionally, it may be helpful to get a hold of someone on the Quova team to learn more about IP geolocation intelligence and its feasibility for our purposes. In particular, I was able to get a hold of Rupert Young, Sr. Director of Software Engineering on the Neustar Data Acquisition & Measurement team and a former Sr. Director of Product Management at Quova. Rupert shared many valuable answers to the questions above and offered to schedule a call with the data science team.

References

[1] Basic ip addressing and troubleshooting guide. Online. Accessed: 2018-06-24.
[2] Dynamic and static ip address differences. Online. Accessed: 2018-06-24.
[3] Geolocation database providers. Online. Accessed: 2018-06-24.
[4] Geolocation software. Online. Accessed: 2018-06-24.
[5] How does an isp assign ip? Online, 2011. Accessed: 2018-06-24.
[6] How does geographic lookup by ip work? Online, 2008. Accessed: 2018-06-24.
[7] How does geolocating by ip address work? Online, 2013. Accessed: 2018-06-24.
[8] Htg explains: Are you using ipv6 yet? should you even care? Online, 2016. Accessed: 2018-06-24.
[9] Number resources. Online. Accessed: 2018-06-24.
[10] Internet society. Online, 2015. Accessed: 2018-06-24.
[11] Oxford university ip address space allocations. Online, 2014. Accessed: 2018-06-24.
[12] W. Parkhurst. Internet addressing and routing first step. Online, 2004. Accessed: 2018-06-24.
[13] Quova. Online. Accessed: 2018-06-24.
[14] Static vs dynamic ip: What do you really need? Online. Accessed: 2018-06-24.
[15] J. Tyson. How internet infrastructure works. Online. Accessed: 2018-06-24.
[16] Why do isps change your ip address? Online. Accessed: 2018-06-24.
[17] Why should you pay more for a static ip address? Online. Accessed: 2018-06-24.
[18] Wikipedia: Ip address assignment. Online. Accessed: 2018-06-24.