A Web Client Perspective on IP Geolocation Accuracy

Joel Sommers
Department of Computer Science
Colgate University
Hamilton, NY 13346, USA
Email: jsommers@colgate.edu

***Abstract*—Geolocation of Internet addresses is widely used by content providers to tailor services and content to users and to restrict access to content. IP geolocation in practice typically relies on databases that include geographic information about individual addresses or address prefixes. Prior studies have assessed the accuracy of these databases by using a set of addresses with known or estimated locations and comparing them with database-reported locations.**

In this paper we investigate IP geolocation accuracy from the standpoint of web clients, by exploiting geolocation information embedded in non-standard HTTP response headers and in unencrypted HTTP cookies. We identify a set of 10,476 websites and content providers that include non-standard HTTP headers and unencrypted cookies with geographic information. We launch HTTP requests to each of these sites from 113 client vantage points with known locations distributed across 6 continents and 60 countries and extract available geographic information from the responses using a battery of hand-crafted regular expressions. We find that the country of the client is included in more than 90% of all responses. Moreover, we observe that about 75% of all responses *only* include the country name or code and that the remaining responses include some combination of geographic information, such as continent, country, city, postcode, region, and coordinates. We observe that accuracy is greatest for the coarsest geographic scope (continent) and least accurate for finer scopes (*e.g.*, coordinate), but that accuracy varies widely across vantage points regardless of the continent or country from which the request is launched.

***Index Terms*—HTTP cookies, HTTP headers, IP geolocation.**

I. Introduction

Many sites and applications on the web today are designed to be *location-aware*. Knowing or estimating the geographic location of a client host can be used for adapting content to a given locale, for restricting content that can or cannot be viewed in a given jurisdiction, for delivering relevant ads, for estimating shipping time or costs, and for many other purposes. Although many smartphones and other devices have embedded GPS radios that can provide accurate geographic location, not all client hosts have such capabilities and not all users want to share their location. To geolocate these hosts, servers typically rely on *geolocation databases* which contain per-address or per-prefix location information, which may include continent, country, city, coordinates, and/or postcode. IP geolocation has been of significant commercial interest for several years as well as an active area of research, and a number of prior studies (*e.g.*, [1]–[10]) have focused on development of new techniques for estimating host locations.

Several past studies have focused on evaluating the accuracy of IP geolocation databases, *e.g.*, [11]–[13]. Evaluating the accuracy of geolocated IP addresses or prefixes poses a significant challenge: it relies on having either precise ground truth locations for a set of hosts/addresses or on ways to estimate the locations of a set of addresses or prefixes with high confidence.

In this paper, we examine IP geolocation accuracy from a new perspective — the viewpoint of web clients. We do this by exploiting geographic information embedded in non-standard HTTP response headers and in unencrypted HTTP cookies. We observe that web application software libraries may cache information about client location in unencrypted cookies and that some sites include client location information in non-standardized HTTP response headers, ostensibly for diagnostics and debugging. By launching HTTP requests from 113 clients with Internet connectivity distributed across six continents and 60 countries with known country—and, for a subset, known city—locations to a set of 10,476 servers that we identify to include geographic information, we gain a new, application-level perspective on IP geolocation accuracy. Although the set of servers we identify to include geographic information is relatively small, it is comprised of hosts that have a broad diversity of ranking within the widely-used Alexa top 1 Million websites. Furthermore, while collecting the geographic data is simple, extracting it poses significant obstacles: there is no standard for how this information is named or structured, and the geographic information extracted may actually refer to the server (or some other location) instead of the client.

We overcome these complications with (1) a manually intensive process of creating regular expressions for identifying geographic data for extraction, and (2) by observing that geographic information in HTTP responses collected by clients at disparate locations should be different if this information is intended to refer to the client's location. Upon extracting and analyzing the geographic information we first observe that the data vary significantly in geographic *scope*, *i.e.*, the

978-1-7281-5628-6/20/$31.00 ©2020 IEEE

level of geographic detail. Perhaps not surprisingly, we also observe that accuracy varies widely across client locations and servers. Specifically, we find that the country of the client is included in more than 90% of all responses and that about 75% of all responses *only* include the country name or code. The remaining responses include some combination of continent, country, city, postcode, region, and coordinates. We observe that accuracy is greatest for the coarsest geographic scope (continent) and least accurate for finer scopes (*e.g.*, coordinate), but that accuracy varies widely across client location and thus network prefix regardless of the continent or country from which the request is launched. Our results provide support for conclusions of prior work regarding inaccuracies in geolocation databases, but from the application-level perspective of web transactions. Our results also provide additional context for prior studies on the effectiveness (or not) of censorship, content blocking or authorization by geolocation [14]–[17].

The contributions of this research are as follows. First, we identify non-standard HTTP response headers and unencrypted HTTP cookies as a new data source for evaluating IP geolocation accuracy and scope. To the best of our knowledge, HTTP responses have not previously been tapped for gaining a perspective on IP geolocation. Second, we develop methods and software for extracting geographic information from HTTP responses. In the interest of openness and for providing ways to reevaluate our work, the software created for this study will be made available to the community. Third, we analyze geographic information embedded in HTTP responses, examining both the scope of geographic information in HTTP responses as well as accuracy.

The rest of this paper is organized as follows. In Section II we discuss related work. In the following section, we describe our client vantage points, how we identify servers to collect measurements from, and our data collection and analysis methods; the results of data analysis are shown in Section IV. Finally, we summarize and discuss future work in Section V.

II. Related Work

An early and experimental Internet standard proposed the inclusion of host coordinates in DNS [18]. Perhaps because of its lack of deployment, geolocation of IP addresses using active measurements or using databases created and provided by analytics firms has been of significant interest over the last several years.

Investigation of techniques for estimating the geographic locations of IP addresses and prefixes has been of longstanding interest. Active measurement-based techniques for estimating host locations generally rely on traceroute and/or latency measurements, typically between *landmarks* with known locations and the hosts for which the location is being estimated [1]–[7], [19], [20]. Besides works that investigate new geolocation techniques, other research has evaluated the geographic span of IP prefixes, finding that associating a prefix with a single location is misleading [21]–[23], and geolocation inaccuracies arising due to NAT and high latencies with mobile devices [24].

Works investigating new active measurement-based techniques have often used hosts with known locations, or free or commercial geolocation databases in order to assess the accuracy of a newly proposed technique. Other works have cast doubt on the accuracy of geolocation databases, *e.g.* [11]–[13], finding that accuracy varies greatly across available databases and that significant inaccuracies exist in all databases. Furthermore, some techniques have used DNS naming hints or WHOIS data for geolocating routers and hosts [25]–[28]. Other work has found that WHOIS records can be the subject of manipulation in order to mask the location of a prefix [16]. The most similar work to ours is that of Guo *et al.* [29] who analyzed web page *contents* for geographic clues to where a server may be located. They used additional measurements such as traceroutes and WHOIS to further refine the estimated server location.

Lastly, a recent work by Weinberg *et al.* [30] uses refined versions of existing active measurement-based geolocation techniques to identify locations of proxies for a number of commercial VPN services in order to evaluate claims made by these VPN providers about where their connection points exist. We use this latter work to validate locations of our VPN-based client locations.

Our work differs from prior work in this area in that we examine IP geolocation from the perspective of web transactions. We do not have information about what geolocation technique (either database or active measurement technique) is used by any server, nor do we attempt in this paper to associate the geographic information identified in HTTP responses with any particular database. Evidence we show in Section IV-B suggests, however, that it may be possible; this is the subject of ongoing and future work.

III. Data Collection and Analysis Methods

There are two steps involved in the collection of HTTP responses and extraction of geolocation data which form the basis of our analysis. First, we identify servers of interest from which to collect HTTP responses. Second, we collect HTTP responses and extract geographic information from them. Our method relies on launching HTTP requests from a set of VPN-based clients with known locations; we also discuss in this section how we validate those locations.

A. Identifying servers of interest

To identify servers that embed geolocation information in HTTP responses, we first aggregated two weeks of Alexa top 1M lists (28 February 2019–13 March 2019) to produce a combined set of 4,625,371 unique domain names. We used the Alexa toplist data available through `toplists.github.io` [31]. From a client attached to our university network, we made HTTP requests to this set of domains, capturing all response headers for each HTTP transaction. For each HTTP request, we followed HTTP redirects up to 30 times before declaring failure. We manually inspected

the response data to develop a set of regular expressions (regexes) to identify a set of hosts that include geographic information in HTTP responses. While the manual inspection of response data was intensive, our only goal at this stage was to identify hosts that *probably* include geographic information. In other words, the set of hosts we identified at this stage surely included false positives (*i.e.*, servers that do not include any geographic information); the elimination of false positives happened at a later step.

Code for this step was written in Go and Python. There are approximately 100 regexes used to detect geographic information, examples of which include `country`, `continent`, `latitude`, `lng`, `geo.{0,2}(ip|loc|code|data|info|cc)`, and `(region|area).*code`. We also used regexes to search for geographic information about our university site in responses, such as the city, region, country names, and coordinate digits. We did this to amass as broad a set of HTTP server domain names as possible that may include geographic information in responses.

The set of servers resulting from this step was 13,553. Although this number is small in comparison to the set of hosts from which we gathered responses, it represents a large variety of ranks in the Alexa top 1M lists, as shown in Figure 1. Note that for hosts that appear in multiple top lists, we used the *lowest* rank observed for constructing the plot. We see in the plot that the overall distribution of hosts is skewed toward higher ranks, but that a very broad set of ranks are included. Moreover, these servers come from 5,632 unique /24 IPv4 prefixes. Of these, about 12% are hosted within Amazon; this was the most of any cloud and/or hosting provider that we were able to identify.

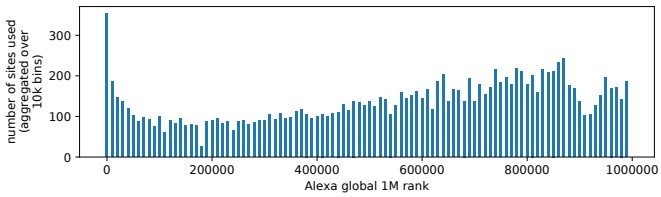

Fig. 1. Alexa top 1M Ranks of hosts that are identified to include geographic information in HTTP responses.

B. Validating client VPN locations

For gathering a set of HTTP responses from which to extract and analyze geolocation information, we used a commercial VPN provider in addition to a host on our university's network. The connection points we chose from the VPN provider are distributed across 6 continents and 60 countries. For a smaller set of client locations, the VPN provider identifies the city location. We requested from the VPN provider to confirm the country location of each connection point and the city locations for those locations that identify the city, which the provider indeed did. Recent work, however, has cast doubt that the locations provided by some VPN providers are accurate [30]. We used the software provided by this work to further validate the locations of these VPN sites; there were no instances of conflicting information. Furthermore, we ran traceroutes through each of these locations back to our university and analyzed DNS names available. Again, in no instance did we find conflicting information, *i.e.*, no location information inferred in a DNS name conflicted with the provider-given location, and in all cases any location-relevant information in DNS names supported the provider-given location. Thus, we have evidence to believe that our clients were connected from the locations where we claim them to be.

We also compared the geographic information about our set of client/VPN locations with publicly available IP geolocation data sources. In particular, we used Maxmind's free GeoIP service [32], ipgeolocation.io [33], ipinfo.io [34], and country-level information available through Team Cymru's WHOIS service [35]. Results are shown in Table II. We observe that continent and country-level information from these databases is in strong agreement with VPN-advertised locations for all sources except for the WHOIS information (which is known to be less accurate and falsifiable [16]). Where the three databases are incorrect, there is little overlap on what locations they are incorrect about. In particular, the only VPN location for which country-level information was incorrect for more than one data source was Algeria (ipgeolocation.io and MaxMind indicate Cyprus, and the WHOIS data indicates the Netherlands; ipinfo.io is correct). Otherwise, there is no overlap in erroneous locations. As we will see later in this paper, the high accuracy for the three databases (ipgeolocation.io, ipinfo.io, and MaxMind) we observe here—specifically for country location—and low accuracy for some locations we observe in our results below suggests that some servers use low-quality geolocation data.

In total, we used 113 client locations (112 from the commercial VPN). There is some redundancy in these locations; in a small number of cases there are multiple connection points located in the same city but attached to a different network provider (and thus with a different IP prefix). In fact, IP addresses for all client locations except for two locations in Singapore are in distinct /24 IPv4 prefixes (i.e., the two Singapore VPN locations use the same /24 prefix). In this study, we only use IPv4-attached locations; we are considering IPv6 locations in our future work. Table I provides a summary of the client locations used in this study.

TABLE I
NUMBER OF CLIENT LOCATIONS ON EACH CONTINENT AND NUMBER OF COUNTRIES INCLUDED.

Continent	AF	AS	EU	NA	OC	SA	Total
Locations	4	19	48	30	7	5	113
Countries	4	12	32	6	2	4	60

C. Extracting geographic information from HTTP responses

The last step of our data collection process was to launch HTTP requests using each of the 13,553 server domain names

TABLE II
FRACTION OF PUBLICLY AVAILABLE GEOLOCATION DATA SOURCES THAT AGREE WITH VPN-ADVERTISED LOCATIONS.

Data source	Continent	Country	Coordinates
Team Cymru WHOIS	—	0.65	—
ipgeolocation.io	0.98	0.96	0.84
ipinfo.io	—	0.96	0.75
MaxMind (Free)	0.98	0.98	0.75

from each of the 113 client locations. We manually created a set of regular expressions and code (in Python) to extract geographic information from the HTTP responses collected at this step. Recall that the first step in our data collection process only identifies hosts that probably include geographic information in HTTP responses; the regexes used at this final step are fine-grained. As a result, we found it necessary to create far more regexes at this step compared with our initial set ($\approx$100 for the first step and $\approx$2000 for this step). As part of our future work, we intend to develop methods for automatically generating regexes to extract geographic information using techniques such as those in [36]–[38].

Table III shows three examples each of unencrypted cookies and non-standard HTTP response headers that contain geographic information about the client (and for which we wrote regexes). We note that it is common practice for web application frameworks and libraries to cache IP geolocation information in cookies (both encrypted and unencrypted) to avoid further queries for that IP address, *e.g.*, see [39], [40], providing at least one reason why such information may appear in cookies. Further, we observed some non-standard header names to explicitly include words such as "debug", suggesting that at least one use for this information is in debugging and diagnostics of site operation. Each of the examples in the table was collected from a client connected through Frankfurt am Main, Germany. These examples are drawn from responses from six different servers.

In the examples in Table III, we first observe different geographic *scopes* referenced, *e.g.*, country, region, city, postcode and/or coordinates. We also observe a wide variety of text structure in the header values. On the one hand, two of the non-standard HTTP response header values are simply ISO 3166 alpha-2 country [41] or continent codes. On the other hand, while other values have recognizable geographic information there is no standard pattern. For example, the first cookie uses a colon character (`:`) as a separator in a string containing geographic information while the other two cookie values use a JSON (or JSON-like) syntax and the third non-standard HTTP header uses other types of separators. Lastly, with the first unencrypted cookie example we also observe some inaccuracy in the city reported (Berlin instead of Frankfurt am Main).

The variety of examples in Table III highlights the challenge of defining regexes to identify geographic information and then parsing the data. For example, for the first cookie in the table, extraction would involve treating the colon as a field separator and splitting the string to derive the country code, region code, etc. This is a "simple" example: we observed some sites to include complex JSON structures and other sites to include serialized PHP objects, among many other peculiarities. Since there are no standards governing how these data are embedded, we manually wrote code to identify and extract each of the relevant pieces of geographic information from a given header or cookie value. In our code we included a number of self-consistency checks in order to minimize the errors that might have been introduced through this process. For example, we compared country and continent names and codes against published lists, *e.g.* [42], and whether coordinate values were within the correct numeric range, among other checks. We also emitted a file containing headers and cookies from which no geographic information was extracted and manually went through these items to make sure that important information was not missed.

We observed that for some sites all extracted geographic information was identical, indicating that the information is about the *server* site (or some other location) rather than the client. We identified all such sites (a total of 3,386) and discarded results for them[1]. In the end, we extracted usable geographic information from a total 10,476 sites. Of the sites and HTTP transactions that we do not discard, there were 590,512 non-standard HTTP headers and 993,791 unencrypted cookies from which we extracted geographic information. Interestingly, the number of bytes we extracted from cookies varies greatly: header values consumed on average 29 bytes (99th percentile is 50 bytes), but cookie values consumed 61 bytes on average, and the distribution was heavily skewed. The median number of bytes in a cookie value was only 24 bytes, but the 90th percentile was 129 bytes, and the 99th percentile was 508 bytes. Looking into the values themselves, we also observed that the number of decimal digits included with coordinates varied and represented highly misleading precision. The mode of the distribution was 4 decimal digits (31253 measurements out of $\approx$ 60000, which represents roughly precision of 11 meters), but there were $\approx$ 15000 measurements that included 5 or more decimal digits (about 1 meter of precision) and some responses included coordinates with 21 digits of precision!

IV. RESULTS

We analyzed two main aspects of the geographic information extracted from the HTTP response data collected and described in the previous section: (1) What is the geographic scope of information that is included with HTTP responses? Does the geographic information in HTTP responses vary from the perspective of different client locations?; and (2) How accurate is the geographic information embedded in HTTP responses compared with known client locations? In this section, we address these questions. Below, we use the

[1]A potential issue with our site filtering method may occur if there are multiple servers at distinct geographic locations which use IP Anycast: location information in HTTP responses about the server will also be distinct but our method may infer those locations to be about the client. From examination of HTTP headers and cookies and BGP routing information from multiple locations, we do not believe that there are servers we used that are affected by this issue.

TABLE III
EXAMPLES OF UNENCRYPTED COOKIES AND HTTP RESPONSE HEADERS CONTAINING GEOGRAPHIC INFORMATION.

unencrypted cookies

```
1  set-cookie: GeoIP=DE:BE:Berlin: ...
2  set-cookie: fly_geo={"countryCode": "de"}; ...
3  set-cookie: geo=j:{"range":[1533874176,1533878271],"country":"DE","region":"",
                      "city":"","ll":[51.2993,9.491],"metro":0,"zip":0}
```

non-standard HTTP response headers

```
1  x-geoip-country-code: DE
2  Geoip_city_continent_code: EU
3  X-Location: lat=51.2993;lng=9.491;country=Germany;city=;ip= ...
```

term *site* to refer to an HTTP server represented by a particular domain name.

A. Scope of geographic information in HTTP responses

We first evaluate the scope or geographic granularity of information extracted from HTTP responses. We consider continent, country, province/state/region, city, postcode, and coordinates (latitude and longitude) as scopes of interest. Figure 2 considers the fraction of all HTTP responses that include information for a given scope. We show results for *all* responses across all client locations, and results broken down by continent (*i.e.*, results for clients within a given continent are aggregated). We first observe that country is—by far—the most commonly-included geographic information; more than 90% of sites across all continents and locations include a country code or name. Overall, there are 7,872 sites that *only* include a country name or code, which is about 75% of all sites. If these results are reflective of the broader ecosystem of websites, they suggest that the majority of sites are concerned with coarse (*i.e.*, country-level) rather than detailed (*e.g.*, coordinate-level) information.

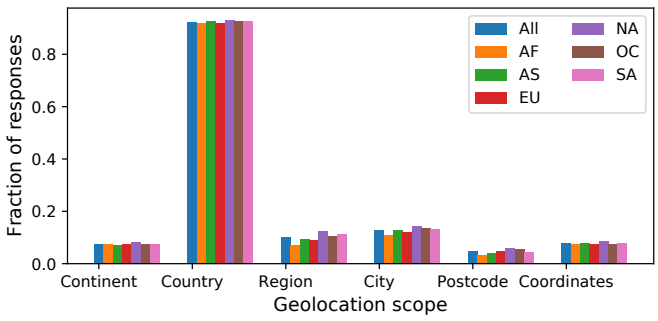

Fig. 2. Fraction of HTTP responses that include information of a given geographic scope.

For the remaining 25% of sites, responses include multiple types of geographic information. The most common combination is to include country, region, and city information (172), but there are many variations beyond that. Overall we observed 32 combinations of different geographic information in responses for a given site. We also analyzed how the scope of geographic information included in a response varies from the perspective of different client locations for a given site. We found that 8,818 sites respond with geographic information of exactly the same scope across all client locations (about 84% of sites) but that the remaining 16% of sites respond differently to requests arriving from different locations. We hypothesize that these differences are due to the availability of geographic information in different databases for different addresses or prefixes, *e.g.*, the postcode and/or coordinates may be available for some locations but not others.

For a number of sites, no geographic information is embedded in the HTTP response when a request is made from a particular client location, suggesting that the geolocation database(s) or other techniques used by a given site are incomplete or failed. We view it as less likely that the results are intentionally omitted, though this is also a possibility. Although we do not show detailed results, we found that VPs for which no geographic information is provided are not concentrated on any one continent. The VP with the most responses that do not include geographic information information is in South Africa; these responses comprise about 5% of all responses. For the majority of VPs, fewer than 1% of responses do not contain any geographic information.

B. Accuracy of geographic information in HTTP responses

To assess the accuracy of the information embedded in HTTP responses, we start with known information about each client location (country, and possibly city). For each country or country/city combination, we queried the Nominatim service of OpenStreetMaps [43] to obtain a coordinate bounding box around the country or country+city of the client and the region name if the city is known. Note that since we consider a coarse bounding box our evaluation of coordinate accuracy is imprecise since areas are almost certainly included in the bounding box that are not actually part of the country or city in question. Still, more than 90% of the bounding boxes we identify for VPs with known city locations are tighter than the 40 km bounds used in prior work [11], [21]. Lastly, we do not show detailed results regarding continent, region (*e.g.*, state or province) or city below due to space limitations.

First, we examine the accuracy of country information included in HTTP responses in Figure 3. Client locations are sorted from least to most accurate along the x-axis. Moreover, only responses from sites that include an indication of country are included. We observe that for about 100 locations, country accuracy is 80% or better. For some locations, however,

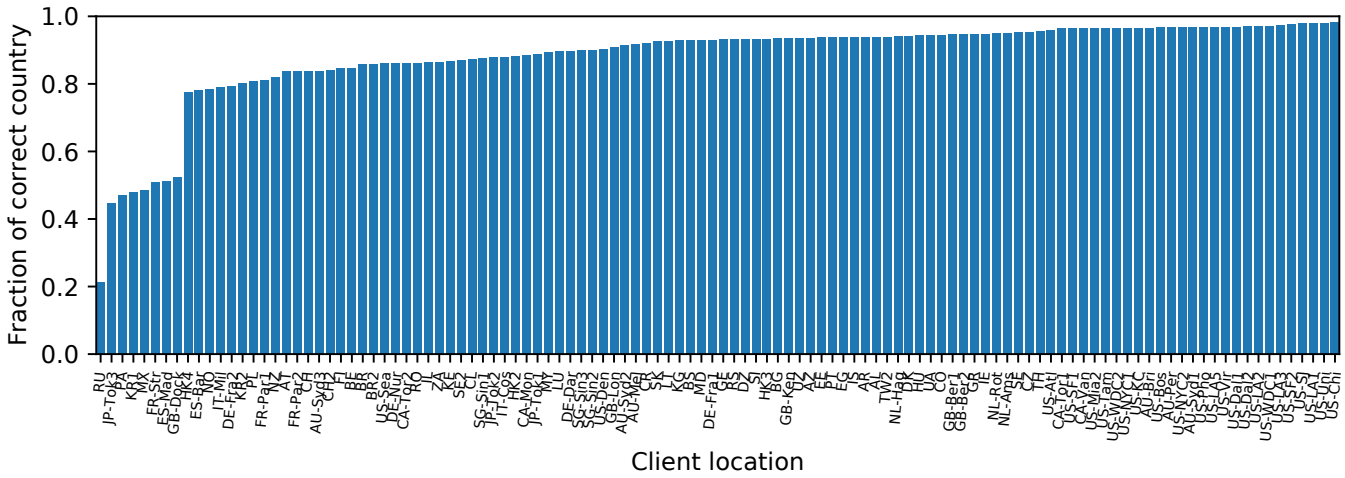

Fig. 3. Accuracy of country information in HTTP responses.

accuracy is poor. For example, with the RU location, about 72% of responses indicate that the site is in Kazakhstan instead of Russia and for the PA location about 38% of responses indicate that the site is in Argentina instead of Panama. Although not shown, we note that for all but 10 locations, continent accuracy is 90% or greater; accuracy for 3 of the remaining locations is less than 60%. Overall, we do not observe any clear relationship regarding client location and country or continent geolocation accuracy. We also do not observe a relationship between RIR of the client prefix and geolocation accuracy (*e.g.*, the 5 least accurate client locations have prefixes in 4 of 5 RIRs). We hypothesize that accuracy is less related to the specific country or continent of the client than the geographic reach of the prefix used; prior work has also suggested this reason for poor geolocation accuracy [21], [22].

In Figure 4 we show accuracy results of geographic coordinates embedded in HTTP responses. We consider the coordinate *accurate* if it is within the bounding box for the country or city of the client. We show plots for country-level accuracy in the top plot and for city-level accuracy in the bottom plot. We first observe a similar profile as results for country code accuracy (*cf.* Fig. 3). In particular, we see that many country-level locations are 80% accurate or better, but that there are a smaller number with lower (or much lower) accuracy. In this plot we only examine VPs for which the country is known. We focus on country identification because it is the granularity typically used for content blocking or authorization.

Interestingly, there are some VPs for which the country or country code indicated in HTTP headers and cookies is rather accurate but for which coordinates are not, and vice versa. In particular, while coordinates for the RU VP are quite accurate, the country information is not, and while coordinates for the GE location are inaccurate, the country code is highly accurate (*cf.*, Fig. 3). This observation suggests possible internal inconsistencies in geolocation databases. For city-level accuracy (bottom plot of Figure 4), we observe that the majority of locations have significant inaccuracies; only a small number are 80% accurate or better.

Figure 5 provides a different perspective on the accuracy of coordinates embedded in HTTP responses. The plots show empirical cumulative distribution functions of distances between coordinates embedded in a response and the center of the true location; distance is computed using the Haversine formula. The figure shows results for locations within Australia (left plot), Canada (center plot), and Spain (right plot); note that these are all city-level locations. We first observe that for some locations (*e.g.*, AU-Syd2, Au-Syd3, AU-Bri, CA-Tor1, and CA-Van) accuracy is high; more than 85% of the coordinates we extract are within 10km of the true location, and certainly within city limits. Accuracy for other client locations, however, can be poor. For the AU-Per location, for example, less than 30% of coordinates are within 10km of the true location, and the remaining coordinates are 1800km away or farther. Interestingly, we also observe from the vertical lines in the figure that there are a number of coordinates extracted from HTTP responses that are identical or very nearly so. We hypothesize that this effect is caused by a relatively small number of geolocation databases or geolocation API providers in use in the Internet. For example, with AU-Per, one may estimate this number at somewhat greater than 4 (there are at least 4 distinct vertical line segments in the curve for AU-Per).

V. Summary and Future Work

In this paper, we evaluate IP geolocation from the perspective of web clients and geographic information embedded in non-standard HTTP response headers and in unencrypted HTTP cookies. We identify a set of 10,476 sites that include geographic information in HTTP responses and launch HTTP requests from a set of 113 client locations with known country or country+city locations. We find that the country of the client is included in more than 90% of all responses, that about 75% of all responses *only* include the country name or code

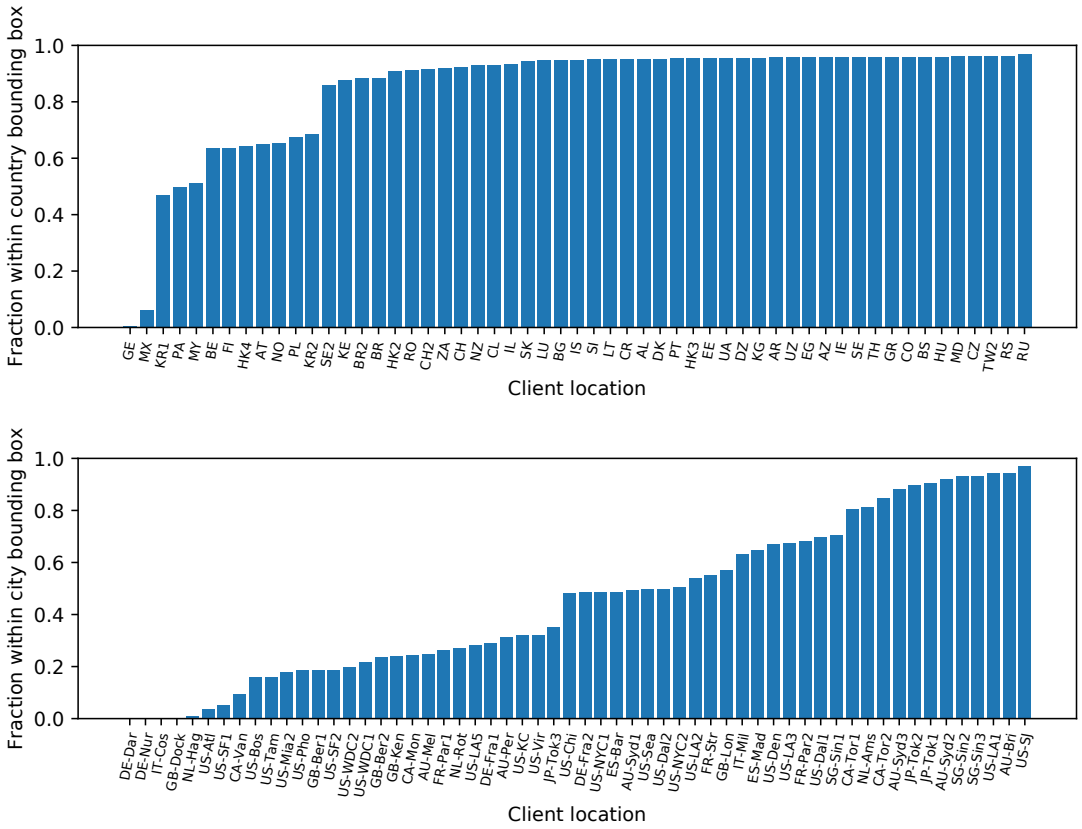

Fig. 4. Accuracy of latitude/longitude coordinates considering if the coordinates lie within a bounding box around the client location. Results for VPs with known country locations are shown on top; results for known city locations are on bottom.

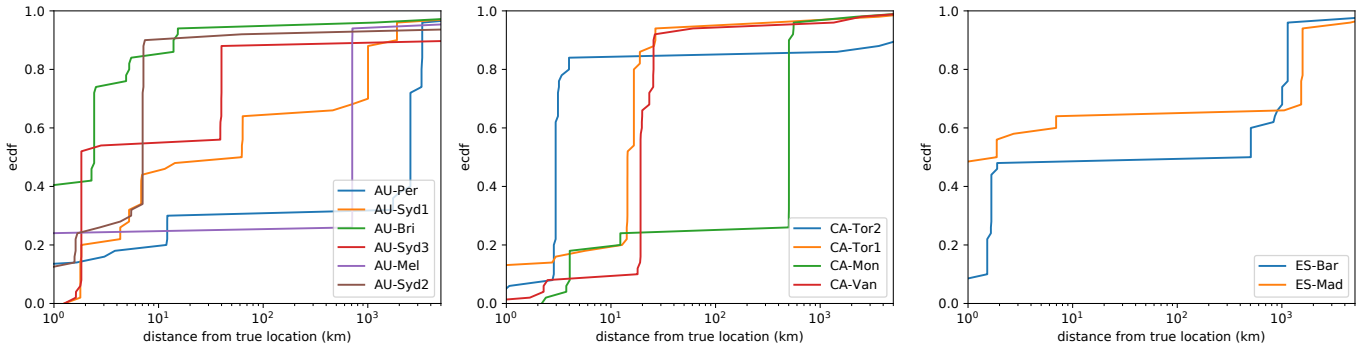

Fig. 5. Empirical CDF of distance from true location for coordinates in HTTP responses for locations in AU (left), CA (center), and ES (right).

and that the remaining responses include some combination of additional geographic information. We observe that accuracy is greatest for the broadest geographic scopes (continent and country) and least accurate for finer scopes, but that accuracy varies widely across client locations and network prefix. In future work, we intend to develop methods for inferring the source of geographic information for a set of websites (*e.g.*, which database or API) and to analyze the effects of inaccurate geolocation on Internet censorship and content authorization.

References

[1] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," *IEEE/ACM Transactions on Networking (TON)*, vol. 14, no. 6, pp. 1219–1232, 2006.

[2] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A learning-based approach for IP geolocation," in *International Conference on Passive and Active Network Measurement*. Springer, 2010, pp. 171–180.

[3] E. Katz-Bassett, J. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. ACM, 2006, pp. 71–84.

[4] Z. Hu, J. Heidemann, and Y. Pradkin, "Towards geolocation of millions of IP addresses," in *Proceedings of the 2012 Internet Measurement Conference*. ACM, 2012, pp. 123–130.

[5] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," in *NSDI*, vol. 11, 2011, pp. 27–27.

[6] B. Eriksson, P. Barford, B. Maggs, and R. Nowak, "Posit: a lightweight approach for IP geolocation," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 2, pp. 2–11, 2012.

[7] B. Wong, I. Stoyanov, and E. Sirer, "Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts," in *NSDI*, vol. 7, 2007, pp. 23–23.

[8] I. Youn, B. Mark, and D. Richards, "Statistical geolocation of Internet hosts," in *2009 Proceedings of 18th International Conference on Computer Communications and Networks*. IEEE, 2009, pp. 1–6.

[9] Z. Dong, R. Perera, R. Chandramouli, and K. Subbalakshmi, "Network measurement based modeling and optimization for IP geolocation," *Computer Networks*, vol. 56, no. 1, pp. 85–98, 2012.

[10] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, and Y. Zhang, "IP-geolocation mapping for moderately connected Internet regions," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 381–391, 2013.

[11] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A look at router geolocation in public and commercial databases," in *Proceedings of the 2017 Internet Measurement Conference*. ACM, 2017, pp. 463–469.

[12] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[13] I. Poese, S. Uhlig, M. Kaafar, B. Donnet, and B. Gueye, "Ip geolocation databases: Unreliable?" *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.

[14] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J. A. Halderman, and R. Ensafi, "403 forbidden: a global view of CDN geoblocking," in *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018, pp. 218–230.

[15] A. Dabrowski, G. Merzdovnik, J. Ullrich, G. Sendera, and E. Weippl, "Measuring Cookies and Web Privacy in a Post-GDPR World," in *International Conference on Passive and Active Network Measurement*. Springer, 2019, pp. 258–270.

[16] J. A. Muir and P. C. V. Oorschot, "Internet geolocation: Evasion and counterevasion," *Acm computing surveys (csur)*, vol. 42, no. 1, p. 4, 2009.

[17] "How to Block Entire Countries from Accessing Your Website," https://www.sitepoint.com/how-to-block-entire-countries-from-accessing-website/, April 2015, accessed January 2020.

[18] C. Davis, P. Vixie, T. Goodwin, and I. Dickinson, "RFC 1876: A Means for Expressing Location Information in the Domain Name System," January 1996.

[19] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *ACM SIGCOMM*, 2001.

[20] H. Maziku, S. Shetty, K. Han, and T. Rogers, "Enhancing the classification accuracy of IP geolocation," in *MILCOM 2012-2012 IEEE Military Communications Conference*. IEEE, 2012, pp. 1–6.

[21] M. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proceedings of the 5th ACM SIGCOMM conference on Internet measurement*, 2005.

[22] B. Gueye, S. Uhlig, and S. Fdida, "Investigating the imprecision of IP block-based geolocation," in *International Conference on Passive and Active Network Measurement*. Springer, 2007, pp. 237–240.

[23] S. Siwpersad, B. Gueye, and S. Uhlig, "Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts," in *International Conference on Passive and Active Network Measurement*. Springer, 2008, pp. 11–20.

[24] M. Balakrishnan, I. Mohomed, and V. Ramasubramanian, "Where's that phone?: geolocating IP addresses on 3G networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. ACM, 2009, pp. 294–300.

[25] Q. Scheitle, O. Gasser, P. Sattler, and G. Carle, "HLOC: Hints-based geolocation leveraging multiple measurement frameworks," in *Network Traffic Measurement and Analysis Conference (TMA), 2017*. IEEE, 2017, pp. 1–9.

[26] B. Huffaker, M. Fomenkov, and kc claffy, "DRoP: DNS-Based Router Positioning," *ACM SIGCOMM Computer Communication Review*, 2014.

[27] P. Endo and D. Sadok, "WHOIS based geolocation: A strategy to geolocate Internet hosts," in *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*. IEEE, 2010, pp. 408–413.

[28] O. Dan, V. Parikh, and B. D. Davison, "IP Geolocation through Reverse DNS," *arXiv preprint arXiv:1811.04288*, 2018.

[29] C. Guo, Y. Liu, W. Shen, H. Wang, Q. Yu, and Y. Zhang, "Mining the web and the Internet for accurate IP address geolocations," in *IEEE INFOCOM 2009*. IEEE, 2009, pp. 2841–2845.

[30] Z. Weinberg, S. Cho, N. Christin, V. Sekar, and P. Gill, "How to Catch when Proxies Lie," in *Proceedings of the 18th ACM SIGCOMM conference on Internet measurement*, 2018.

[31] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. Strowes, and N. Vallina-Rodriguez, "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists," in *Proceedings of the 2018 Internet Measurement Conference*, 2018.

[32] "MaxMind: GeoLite2 Free Downloadable Databases," https://dev.maxmind.com/geoip/geoip2/geolite2/, accessed January 2020.

[33] "ipgeolocation: Free IP Geolocation API and IP Location Lookup Database," https://ipgeolocation.io, accessed January 2020.

[34] "ipinfo.io: The Trusted Source for IP Address Data," https://ipinfo.io, accessed January 2020.

[35] "Team Cymru IP to ASN Lookup v1.0," https://whois.cymru.com, accessed January 2020.

[36] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Jagadish, "Regular expression learning for information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 21–30.

[37] H. Fernau, "Algorithms for learning regular expressions from positive data," *Information and Computation*, vol. 207, no. 4, pp. 521–541, 2009.

[38] M. Luckie, B. Huffaker *et al.*, "Learning regexes to extract router names from hostnames," in *Proceedings of the Internet Measurement Conference*. ACM, 2019, pp. 337–350.

[39] "Geocoder," https://github.com/alexreisner/geocoder, accessed January 2020.

[40] "Geokit-rails," https://github.com/geokit/geokit-rails, accessed January 2020.

[41] "International Organization for Standardization Country Codes — ISO 3166," https://www.iso.org/iso-3166-country-codes.html, accessed January 2020.

[42] "Comprehensive country codes: ISO 3166, ITU, ISO 4217 currency codes and many more," https://datahub.io/core/country-codes, accessed January 2020.

[43] "Nominatim," https://nominatim.openstreetmap.org/, accessed January 2020.