Assessing Co-Locality of IP Blocks

Manaf Gharaibeh, Han Zhang, Christos Papadopoulos
Colorado State University
Fort Collins, CO, USA
Email: gharaibe, zhang, christos@cs.colostate.edu

John Heidemann
University of Southern California/ISI
Los Angeles, CA, USA
Email: johnh@isi.edu

***Abstract*—Many IP Geolocation services and applications assume that all IP addresses within the same /24 IPv4 prefix (a */24 block*) reside in close physical proximity. For blocks that contain addresses in very different locations (such as blocks identifying network backbones), this assumption can result in a large geolocation error. In this paper we evaluate the co-location assumption. We first develop and validate a hierarchical clustering method to find clusters of IP addresses with similar observed delay measurements within /24 blocks. We validate our methodology against two ground-truth datasets, confirming that 93% of the identified multi-cluster blocks are true positives with multiple physical locations and an upper bound for false positives of only about 5.4%. We then apply our methodology to a large dataset of 1.41M /24 blocks extracted from a delay-measurement study of the entire responsive IPv4 address space. We find that about 247K (17%) out of 1.41M blocks are not co-located, thus quantifying the error in the /24 block co-location assumption.**

I. Introduction

Internet location-aware applications and research in geolocation benefit from IP-to-geolocation provided by services such as MaxMind [4], IP2Location [3], and DB-IP [1]. These services provide various degrees of geolocation accuracy. While the accuracy has improved over the last decade [21], [20], [14], accurate (city or street level) IP-geolocation is still regarded as an open problem. Achieving good accuracy can be difficult when some blocks of adjacent IP addresses span large geographic areas [18], [10]. Today, most services assume that the addresses within the same /24 prefix (a */24 block*) are geographically proximate—the *block co-locality assumption*. When this assumption is violated, some addresses in the block will have poor geolocation accuracy.

Public databases such as MaxMind *GeoLiteCity* [4] and IP2Location *LITE-DB11* [3] assume block co-locality. The entries in these databases identify IP blocks of various sizes and assign each a specific location. MaxMind's database contains about 1.7M such blocks covering 3.6B addresses (97% of the allocated address space). The IP2Location database has 2.2M entries covering the entire IPv4 address space (no locations are assigned to special blocks such as multicast). In these databases nearly all blocks are /24 or larger, thus 99% or more of the /24 blocks are marked as co-located. Some location-aware applications also seem to adhere to the /24 co-locality assumption. An architecture proposed by Chen et al. [7] maps a client's request to a proximal content server based on prefixes, meaning that all clients within the same prefix are mapped to the same content server. They suggest the mapping at /20 prefix granularity to minimize the number of required mappings.

In this paper we assess the /24 block co-locality assumption. We leverage a large subset of the dataset collected by Hu et al. [13] and publicly available [6]. The dataset contains round-trip estimates for every responsive address in the IPv4 address space measured from several vantage points (VPs). We assess blocks co-locality in this dataset based on the observation that geographically co-located hosts will show similar network delays when probed by the same set of VPs [16]. Based on this observation, we cluster addresses in each /24 block into groups by similarity of the delay measurements from multiple VPs. We then identify /24 blocks with multiple clusters and show that these clusters violate the block co-locality assumption and likely contain addresses in distinct locations.

Our first contribution is to introduce and evaluate a methodology to assess co-locality of endpoints in an IP block. In this paper we limit our study to /24 blocks, but our methodology is independent of block size. Our delay-based clustering algorithm automatically identifies blocks that appear to have endpoints at different locations. We validate the accuracy of this method against two ground truth datasets (Section IV): first, a set of /24 blocks that we have selected based on our belief that they are co-located, and second, an artificially-constructed set of multi-location blocks. We confirm that 93% of the blocks identified as multi-location blocks in the ground truth datasets are true positives. Our second contribution is the application of this methodology to analyze 1.41M /24 blocks (118M addresses), a large part of the Internet. We find that a noticeable fraction of these blocks (17%, or 247K blocks) appear to have endpoints at multiple locations.

II. Dataset Description

Our analysis uses the IP geolocation dataset collected by researchers at ISI [6] extended from prior work by Hu et al. [13]. The original dataset contains round-trip time measurements for all the allocated and responsive IP addresses in the IPv4 address space. The dataset has about 472M IP addresses in just under 3.5M /24 blocks and was collected from Feb. 2012 to Mar. 2013. RTTs were measured from about 670 vantage points (VPs) on PlanetLab. The work used the following algorithm to pick the 10 closest VPs to any /24 block. First, all available VPs were used to probe a few representative IP addresses in a block. The 10 VPs with shortest RTTs were selected to probe all IP addresses in that block. The use of

VPs close to the target minimizes interference from congestion and maximizes the precision of geolocation (something 400 milliseconds away can be anywhere on earth, but something within few milliseconds is likely in the same city). Also to reduce congestion noise, latency was reported as the *minimum* of 10 measurements. For our work, we used the raw probing data for all /24 blocks where each block contained at least 10 IP addresses that responded to *all* VPs probes. The delay measurements of each IP address are treated as its coordinates in a multidimensional space. Our dataset comprises of **118.5M** IP addresses in **1.41M** /24 blocks.

III. Methodology

A. Identifying Multi-Location Blocks

Our methodology is based on the insight that geographically co-located IP addresses from the same IP block exhibit relatively similar network delays when probed from the same vantage points (VPs). For each IP address we create a vector of 10 delay measurements observed from 10 different VPs (Section II). In order to identify groups of likely co-located IP addresses we formulate the problem as finding similar IP addresses in a multidimensional space of delay coordinates. Co-located IP addresses are expected to have small distances between them in the delay multidimensional space. So we cluster IP addresses in a block based on the similarity of their delay vectors. A block with all of its IP addresses mapped into one cluster is likely a single-location block, while a block with 2 or more clusters is likely a multi-location block.

To identify clusters of co-located IP addresses in a block we use an agglomerative hierarchical clustering algorithm from **R** *cluster* package called *agnes*. Given the delay vectors of the IP addresses in a block, the algorithm generates a hierarchical structure (dendrogram) based on the dissimilarities of the delay vectors. We use the *Standardized Euclidean* distance metric to measure dissimilarities. We use a dynamic tree cut method from the *dynamicTreeCut* package [15] to identify the clusters in the dendrogram. The combination of the these methods satisfies the need to identify clusters automatically without a prior knowledge of their number or size.

As with other agglomerative hierarchical methods, the *agnes* method generates a bottom-up hierarchical structure for the input observations. Each observation starts as a cluster by itself. In each subsequent step the closest two clusters not already in the same cluster are merged into one larger cluster. The process continues until there is only one cluster of all observations. The height at which two clusters are merged in the tree-like dendrogram structure is computed as a function of the dissimilarity between the two merged clusters. The dissimilarity between two clusters can be computed in different ways. In this work we use the *average linkage* method, which computes the distance between two clusters as the average of pairwise dissimilarities between the objects in the two clusters. For two clusters, say cluster A with n_a objects and cluster B with n_b objects, the metric is computed using Equation 1, where D is the distance metric used to compute the distance between two objects. We use the *Standardized Euclidean* distance metric to balance the depth of the measurements observed from VPs at different distances from targets.

$$d_{average}(A,B) = \frac{1}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} D(IP_{Ai}, IP_{Bj}) \tag{1}$$

Prior work has shown the need for selecting clustering thresholds dynamically when examining Internet RTT data [9]. To identify clusters automatically for each of our 1.41M /24 blocks, we use the "*Dynamic Hybrid*" tree cut method [15] to dynamically identify clusters in a dendrogram. This method uses dendrogram-merging information to build the clusters in a bottom-up fashion. We tuned the method parameters to be conservative on what is considered a cluster. Many of the parameters are set as a fraction of the joining heights of the branches in the dendrogram. The one parameter we found most effective is the *minimum gap* parameter, which specifies the minimum joining height to allow two clusters to be merged. Higher settings of this parameter allow more clusters to be merged. The result is fewer clusters with significant dissimilarities indicating higher probability of being at different locations. We also set the minimum cluster size to 10 to reduce the possibility of getting small clusters of outliers. We evaluate the clustering method using two ground truth datasets in Section IV.

B. Methodology Limitations

Our methodology has the following limitations. First, as with all delay-based methods, our approach can be affected by measurement inaccuracy. This problem is alleviated by taking multiple measurements over time, the use of multiple VPs per block and picking the minimum RTT. Second, our methodology does not identify the geographic locations of the clusters and the actual distance between them. Fortunately, geographical locations are not required to determine if two IP addresses are *co*-located as we propose in our methodology.

IV. Validating Identification of Multi-Location Blocks

In this section we validate our methodology by showing that it accurately finds single- and multi-location blocks in our ground-truth dataset. We build our ground truth dataset as follows. First, we identify single-location blocks as described in Section IV-A. Second, we use this data to construct a multi-location ground truth dataset as described in Section IV-B. Third, we use the constructed ground truth dataset to validate our methodology as described in Section IV-C. Finally, we estimate an upper bound of false positives for the clustering method as described in Section IV-D.

A. Building a Single-Location Ground Truth Dataset

We build a dataset of /24 blocks that we strongly believe are single-location blocks for two purposes: (a) to evaluate the clustering method accuracy on single-location blocks, (b) to build the multi-location ground truth dataset.

Our single-location ground truth dataset is composed of address blocks belonging to selected academic institutions. We opted for academic institutions because they typically have specific, well-defined physical locations with many end-user computers within a small geographical area. Such institutions often host their own web services [20]. Academic institutions are also relatively easy to find on the map and services such as Google Maps already have campus outlines and geographic coordinates for them. Finally, academic institutions tend to be long-lived with more accurate *whois* entries than average. These properties make academic blocks attractive candidates for our purposes.

We begin by identifying the locations and /24 IP address blocks containing the websites of 4650 universities from different locations around the world listed at [5]. We verify that these blocks are locally hosted at their universities by applying two filters. First, we detect outsourcing using *whois* information and discard outsourced blocks. We identify outsourcing by matching the *OrgName* field from *whois* information with the institution name. For example, Duke University's website is at the IP address (54.191.241.8), which the *whois* OrgName identifies as *Amazon Technologies Inc*, this shows evidence of outsourcing and is therefore removed from our list. Second, we use the *Google Maps Geocoding API* [2] to identify a university's physical location (latitude/longitude), and compare this to the physical location assigned to the IP address by MaxMind. We discarded IP addresses where the great circle geographic distance between these geographic locations is more than 10 miles. This step discarded blocks that are located to differing or uncertain locations. The filters described above are very strict and resulted in rejecting otherwise viable entries in our dataset. However, this only increased the confidence of the remaining entries since they passed a higher bar. Finally, we cross-checked the remaining blocks from the university dataset with the one we extracted from the 1.41M blocks and formed our single-location dataset from blocks that appear in both.

This aggressive filtering reduced our initial set of 4650 academic institutions to 85, but resulted in a high-confidence, single-location ground truth dataset of 85 /24 address blocks. We used this dataset to build a multi-location ground truth dataset as described next.

B. Building a Multi-Location Ground-Truth Dataset

We built a multi-location ground truth dataset by combining two single-location blocks to form synthetic multi-location blocks. To generate synthetic multi-location blocks we find all blocks from the single-location dataset that were probed by the same set of VPs; we call these *VP-compatible blocks*. We then computed all two-block combinations in each set of VP-compatible blocks, combining all measurement data from the two blocks to create a new synthetic block. Merged blocks may have up to 512 addresses; however, since we had data from ping-responsive addresses only, merged blocks almost always had fewer addresses, often less than 256. These synthetic blocks form our ground truth multi-location blocks dataset.

Some of our VP-compatible single-location blocks that we use to build the multi-location dataset are actually quite close to each other. We therefore identify two subsets in our multi-location dataset: those composed of *almost-co-located* blocks (within 10 miles of each other), and *not-co-located* blocks (22 miles apart or more). We identify 21 almost-co-located synthetic blocks and 99 not-co-located blocks.

C. Validation

The two ground truth datasets (single- and multi-location) let us evaluate the ability of our delay-based clustering method to identify co-located blocks and blocks that span multiple geographic locations. We tested our method on both single- and multi-location datasets.

We first considered our single-location dataset. Our clustering algorithm classified correctly 91% (77 of 85) of the blocks in our single-location dataset. Seven blocks were identified to have 2 clusters while one block was not clustered. Manual investigation of the 7 blocks with 2 clusters showed distinct latency distributions for the two clusters. On the other hand host name and *traceroute* results did not show any evidence that the IP addresses in any of these blocks were at different locations. We still believe that these blocks were co-located, but some addresses experienced different delays possibly due to a wireless network or a slow switch/router. It is well-known for example, that a wireless access point adds a few milliseconds to the RTT of wireless hosts. We did not investigate these cases further.

We next turn to our synthetic, multi-location dataset. First, we discarded synthetic blocks built from the 7 misclassified single-location blocks (see above) since we know those would be identified as multi-location blocks. We then applied our clustering methodology to the remaining not-co-located 99 synthetic blocks. Fig. 1 shows the number of identified clusters and the corresponding distance between the two combined blocks for each synthetic block. We correctly identified 88% of these as multi-location blocks. Manually investigating the remaining 12% false negatives we saw very similar delay measurements for IP addresses in the combined blocks leading to incorrect identification. Such delay measurements could be a result of a relatively close proximity between the synthetic blocks. Other possible reasons are blocks sharing most of the network hops to the VPs or similar path distances from the VPs. For example, the VP at Berlin Institute of Technology observed similar delay measurements to the two synthetic blocks at the University of Gttingen and Jade University of Applied Sciences in Germany. The VP at Hamburg University of Applied Sciences also observed similar delay measurements to these two blocks. This is one case where the clustering method falsely identified the two blocks as co-located.

To examine the most challenging blocks we also looked at the 21 proximal almost-co-located synthetic blocks where the real-world distance between each block is within 10 miles. One example is the combination of Dongbei University of Finance And Economics and Dalian University of Technology in China, which are less than one mile away from each other.

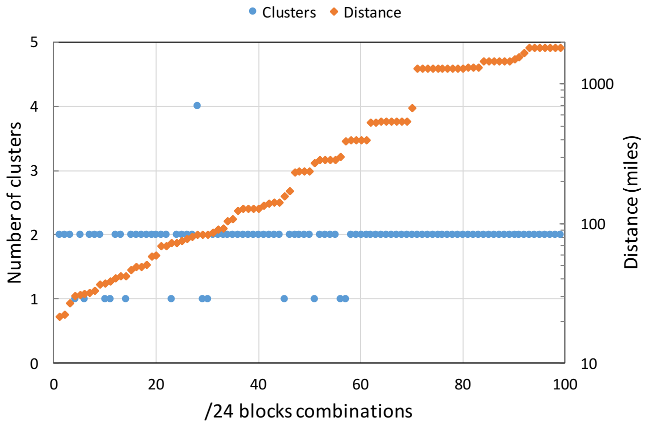

Fig. 1: Results of applying the delay-based clustering method to 99 2-block combinations. The graph shows the number of reported clusters for each synthetic block and the corresponding distance between the combined blocks.

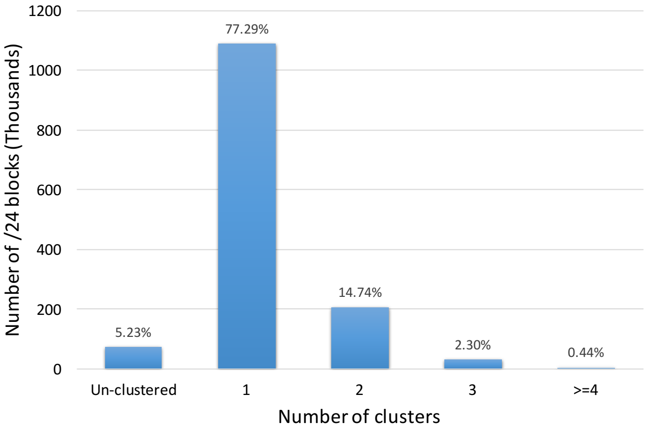

Fig. 2: Distribution of the number of clusters for all 1.41M /24 blocks. More than 17% (∼247K blocks) are identified as multi-location blocks.

Despite close physical proximity, our methodology correctly identifies 38% (8 of 21) of these blocks as multi-location. Overall, 93% of the cases identified as multi-cluster blocks are true positives in our ground truth datasets, which gave us confidence that our methodology works reasonably well.

D. Bounding the False Positives

It is important for our clustering method to maintain a low false-positive rate to ensure that we do not overestimate the number of multi-location blocks (false positives). To estimate an upper bound for the false positive rate, we built another extended set of /24 blocks that are likely co-located. We leveraged again address blocks in academic institutions. We followed a similar procedure as with the single-location dataset, but now we picked a random set of universities from the original set of 4650 institutions that have at least a /16 block assigned to them. As before, we verified they do not include web hosting services. Of these 100 /16 blocks we found 3,062 /24 blocks that appear in the ISI dataset. We ran our clustering method on all these blocks. The results show that 239 blocks (7.8%) are not clustered, 2657 blocks (86.77%) have one cluster, and 166 blocks (5.4%) have 2 clusters. The blocks that failed to cluster did not meet the clustering criteria such as the minimum number of addresses in a cluster. Since any of the blocks identified as multi-location could indeed be multi-location blocks, we regard the 5.4% as an approximate upper-bound for the false positive rate in our clustering method.

V. Co-Locality of /24 Blocks

A. Identifying Multi-Location /24 Blocks

In this section we discuss the results of applying our clustering methodology to the entire ISI dataset. Fig. 2 shows the distribution of clusters in all 1.41M /24 blocks. About 17% (∼247K blocks) appear to have endpoints at multiple locations. 82% of the multi-cluster blocks are grouped into 2 clusters of IP addresses. A small fraction, 0.44%, of the multi-location blocks are grouped into four or more clusters. Our method failed to cluster 73792 /24 blocks (5.23% of all blocks), 98% of which have 20 IP addresses or less. A block is identified as not clustered when the clustering method can not find any cluster with the required minimum number of IP addresses that satisfies all clustering criteria. While this is more typical in blocks with small number of IP addresses, it can also be true for any block with endpoints that are highly scattered geographically.

B. Characterizing Multi-Location Blocks

Our method identified about 247K blocks as multi-location blocks. We found multi-location blocks in 182 different countries in our dataset. We list the top 10 countries sorted by the number of /24 blocks (second column) found in our dataset of 1.41M blocks in Table I. The third and forth columns respectively list multi-location /24 blocks identified per country both as an absolute number and as a percentage of a country total /24 blocks in the dataset. We note different percentages of identified multi-location blocks across different countries. For example, Germany has 42.34% of its blocks in the dataset identified as multi-location. At the other end, only 1.54% of the /24 blocks in China are labeled as multi-location. The fifth column lists each value in column 2 as a percentage of the total blocks assigned to each country by the corresponding RIR. From the table we can see our dataset has a reasonable representation of the IPv4 space for the top 10 countries, ranging from about 7% to over 21%. Differences in multi-location percentages across countries may be due to different policies of IP address assignment. We find that a significant portion of the identified multi-location blocks belong to big ISPs in countries with rich Internet infrastructure such as the United States and Western Europe.

We list the top 10 ISPs (first column) and their ASNs (second column) sorted by the number of /24 blocks (third column) in Table II. The fourth column lists the total number of blocks for each ISP in the dataset while the fifth column is simply the ratio of the previous two columns. From the table we can see that some ISPs (such as DTAG and ONC NTT) show a high percentage of multi-location blocks, while others have much smaller percentages. This difference may reflect different IP allocation policies across ISPs with respect to the geographic distribution of the addresses. It is also worth mentioning that some ISPs dominate the multi-location blocks in their countries in our dataset. For all the countries listed in Table II, more than 40% of their multi-location blocks are from few (less than four) ISPs. We leave further investigation of these phenomena as future work. The sixth column in the table lists each ISP blocks in the dataset (column 4) as a percentage of the total blocks in its ASN announced prefixes (computed based on data from http://cyclops.cs.ucla.edu/). From column 6 We can see that our dataset has a reasonable representation for the top 10 ISPs ranging from about 12% to 49%.

TABLE I: Top 10 countries sorted on their total number of /24 blocks in the dataset and the corresponding multi-location blocks percentage per country.

Country	Blocks in Dataset	Multi-location Blocks	Multi-location Blocks %	% of Country Blocks in Dataset
US	430947	84649	19.64%	6.83%
CN	98016	1507	1.54%	7.45%
DE	81925	34691	42.34%	17.45%
JP	71131	20899	29.38%	8.97%
GB	63609	12339	19.40%	13.24%
KR	60265	10296	17.08%	13.73%
FR	55870	6900	12.35%	17.97%
BR	53050	4772	9.00%	16.57%
RU	37772	2954	7.82%	21.10%
CA	35816	3414	9.53%	12.57%

TABLE II: Top 10 ISPs sorted on their number of multi-location blocks, and their corresponding percentages of ISPs' total number of blocks in the dataset.

ISP Name	ASN	ISP Multi-location Blocks	ISP Blocks in Dataset	Multi-location Blocks %	% ISP Blocks in Dataset	Country
DTAG Deutsche Telekom AG	3320	21204	36359	58.3%	25.5%	DE
COMCAST-7922 - Cable Comm	7922	11804	69117	17.1%	24.2%	US
OCN NTT Comm Corp.	4713	9204	14841	62.0%	12.9%	JP
ATT-INTERNET4 - AT&T Serv	7018	8994	43656	20.6%	11.8%	US
Uninet S.A. de C.V.	8151	7033	24269	29.0%	49.0%	MX
UUNET - MCI Comm Serv	701	6881	28497	24.2%	14.7%	US
CENTURYLINK-US-LEGACY-QWEST	209	6766	26197	25.8%	38.8%	US
BSKYB-BROADBAND-AS Sky UK Ltd	5607	5810	10501	55.3%	40.7%	GB
VODANET Vodafone GmbH	3209	5665	14049	40.3%	41.5%	DE
TPNET Orange Polska Spolka	5617	5561	14950	37.2%	46.6%	PL

VI. Related Work

Much of the prior work in the area of IP geolocation focuses on improving geolocation accuracy [12], [14], [21], [20], [8]. While prior approaches use different techniques, they are mostly delay-based. These approaches are often evaluated on a small number of targets (in the order of a few hundreds). Our work is different in that it does not propose a new algorithm to improve geolocation and is not limited to a small set of targets. Instead, we characterize co-locality of about 1.41M /24 blocks showing that many appear to have endpoints at different physical locations.

Other IP geolocation work studied the accuracy and granularity of public and commercial databases. Poese et al. [17] found that some databases split ISP blocks into smaller ones for more accuracy; however, that made their geolocation accuracy worse. Siwpersad et al. [18] study the geographic resolution of geolocation databases. They compared location information provided by the databases with locations computed using Constraint-Based Geolocation (CBG) [12]. They concluded that the resolution of the databases is way coarser in comparison. Gueye et al. [11] also used CBG to estimate the max distance between block endpoints to estimate its geographic span, which they concluded could be large. Overall, prior work is concerned with the accuracy and granularity of geolocation databases. Our work focused on studying the co-locality of /24 blocks and did not address actual geolocation. We used a different methodology that enabled automatic identification of groups of co-located IP addresses. Compared to other work, our dataset is much larger and representative as well as more recent.

Freedman et al. [10] studied the geographic characteristics if IP prefixes and their influence on BGP routing tables. Their results showed about 1.4% of /24 blocks or smaller span distances of more than 100 miles. They extracted the locations of IP addresses based on DNS naming heuristics using the *undns* tool [19]. DNS IP to location mapping has many shortcomings and can be unreliable due to the lack of naming standards. Our method is not dependent on identifying IP addresses locations; instead we studied IP address proximity using observed latency measurements from multiple VPs.

Fan et al. [9] studied the dynamics of mapping users to Front End (FE) clusters, which are groups of geographically proximate content servers used by Content Distribution Networks (CDNs). They enumerated CDNs' FE servers and then used a clustering technique similar to ours to group the servers into FE clusters. While both methods used similar delay-based clustering techniques, our study has a different purpose. We use clustering to study co-locality of address blocks as opposed to their goal of identifying an FE cluster of one CDN.

Compared to Hu et al. [13], from which we leveraged our dataset, the authors implemented a method to scale existing delay-based geolocation approaches such as Shortest Ping and CBG to geolocate all responsive IPv4 address space. They showed that careful selection of a small number of VPs can maintain a comparable level of accuracy to that of using tens of VPs. While we used a large subset of their raw probing dataset, the problem we addressed is different. Their work used delay measurements to geolocate IP addresses, while we used them as signatures to identify groups of similar endpoints in a block.

VII. Conclusions

Our work introduced a simple clustering methodology to assess IP address block co-locality. We identified groups of IP addresses that appear to be at different locations based on their delay measurements observed from a number of vantage points (VPs). We used a large dataset of 1.41M /24 blocks and showed that more than 17% appear to be multi-location blocks. This outcome disagrees with the common assumption of /24 block co-locality in many geolocation databases. We also found that the majority of the blocks identified as multi-location belong to large ISPs in countries with rich Internet connectivity such as the United States and Western Europe.

An important question that we did not address is how do our results impact the information in geolocation databases such as MaxMind. While it may appear straightforward, such comparison is actually quite hard. Information in databases such as MaxMind is not always accurate, so when disagreements appear it is not clear who is right. Moreover, blocks with different locations in MaxMind may still be close enough to be missed by our methodology. Finally, MaxMind may contain ambiguous information. For example, some /24 blocks were found to have different locations, with part of the block labeled with a city granularity, and the rest with a country granularity. In order to address such questions one needs to build an active measurement infrastructure and perform measurements when there is disagreement.

Our work can be part of a system to evaluate geolocation databases such as MaxMind. Such a system would likely deploy a long-lived active measurement infrastructure and compare results with both free and commercial geolocation databases. A long-lived system can also track the movement of IP address blocks as they get traded. We plan to pursue this as part of future work.

Acknowledgment

The work in this paper is partially sponsored by the Department of Homeland Security (DHS) Science and Technology Directorate, HSARPA, Cyber Security Division, via SPAWAR Systems Center Pacific under Contract No. N66001-13-C-3001 (all authors), and via BAA 11-01-RIKA and Air Force Research Laboratory, Information Directorate under agreement numbers FA8750-12-2-0344 and FA8750-15-2-0224 (Papadopoulos and Heidemann). The U.S. Government is authorized to make reprints for Governmental purposes notwithstanding any copyright.The views contained herein are those of the authors and do not necessarily represent those of DHS or the U.S. Government.

References

[1] The DB-IP database. https://db-ip.com, 2015.

[2] The Google maps geocoding API. https://developers.google.com/maps/documentation/geocoding/, August 2015.

[3] IP2Location. http://www.ip2location.com, 2015.

[4] Maxmind, inc. https://www.maxmind.com, 2015.

[5] Universities worldwide. http://univ.cc, July 2015.

[6] USC/LANDER project. Internet addresses geolocation dataset, PREDICT. https://ant.isi.edu/datasets/geolocation/, 2015.

[7] F. Chen, R. K. Sitaraman, and M. Torres. End-user mapping: Next generation request routing for content delivery. In *The ACM Conference on Special Interest Group on Data Communication*, pages 167–181, 2015.

[8] B. Eriksson, P. Barford, B. Maggs, and R. Nowak. Posit: a lightweight approach for IP geolocation. *SIGMETRICS Perform. Eval. Rev.*, 40(2):2–11, Oct. 2012.

[9] X. Fan, E. Katz-Bassett, and J. S. Heidemann. Assessing affinity between users and CDN sites. TMA, 95-110(2015), 2015.

[10] M. J. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan. Geographic locality of IP prefixes. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, IMC '05, pages 153–158, Berkeley, CA, USA, 2005. USENIX Association.

[11] B. Gueye, S. Uhlig, and S. Fdida. Investigating the imprecision of IP block-based geolocation. In *Proceedings of the 8th International Conference on Passive and Active Network Measurement*, PAM'07, pages 237–240, Berlin, Heidelberg, 2007. Springer-Verlag.

[12] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of Internet hosts. *IEEE/ACM Trans. Netw.*, 14(6):1219–1232, Dec. 2006.

[13] Z. Hu, J. Heidemann, and Y. Pradkin. Towards geolocation of millions of IP addresses. In *The 2012 ACM conference on Internet measurement conference*, IMC '12, pages 123–130, 2012.

[14] E. Katz-Bassett, J. P. 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*, 2006.

[15] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics*, 24(2):719–720, 2008.

[16] V. N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for Internet hosts. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, 2001.

[17] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. Ip geolocation databases: unreliable? *SIGCOMM Comput. Commun. Rev.*, 41(2):53–56, Apr. 2011.

[18] S. S. Siwpersad, B. Gueye, and S. Uhlig. Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts. In *Proceedings of the 9th international conference on Passive and active network measurement*, 2008.

[19] N. Spring, R. Mahajan, and T. Anderson. The causes of path inflation. In *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '03, pages 113–124, New York, NY, USA, 2003. ACM.

[20] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang. Towards street-level client-independent IP geolocation. In *Proceedings of the 8th USENIX conference on Networked systems design and implementation*, 2011.

[21] B. Wong, I. Stoyanov, and E. G. Sirer. Octant: a comprehensive framework for the geolocalization of Internet hosts. In *The 4th USENIX conference on Networked systems design & implementation*, 2007.