Geo-locating BGP prefixes

Philipp Winter
CAIDA, UC San Diego

Ramakrishna Padmanabhan
CAIDA, UC San Diego

Alistair King
CAIDA, UC San Diego

Alberto Dainotti
CAIDA, UC San Diego

***Abstract*—Geo-locating BGP prefixes can help us understand routing anomalies, prefix aggregation, or reveal what regions are affected by an Internet outage. Our work shows that the naive approach to prefix geo-location—simply mapping each IP address to its corresponding geo-location—can be ambiguous because a prefix may contain another, separately-announced prefix that maps to a different geographical location. Should the containing prefix also map to the locations of the contained prefix? We show that this question is difficult to answer and characterize the scope of these ambiguities by geo-locating around 680,000 prefixes to countries, regions, and cities using both GeoLite and NetAcuity Edge. We find that 0.3% of prefixes are ambiguous with respect to countries but these prefixes constitute 8.5% of the IPv4 address space. In the second part of our work, we study the mappings from prefix to location. We find that most prefixes map to only a single city but the shorter a prefix, the more locations it maps to. Our dataset however contains outliers, e.g., a /23 that maps to as many as 127 (potentially spoofed) countries. Our work takes a first look at prefix geo-location and identifies issues one should be aware of, which paves the way towards more sophisticated applications such as the geo-location of autonomous systems. We make our code and datasets publicly available to facilitate further analysis.**

***Index Terms*—BGP, prefix, geolocation**

I. Introduction

The mapping of individual IP addresses to their geographical locations—*i.e.,* IP geo-location—is a well-studied problem and research area [4], [7], [19], [11]. This is not the case for BGP-announced prefixes. A sound approach to geo-locate BGP prefixes would improve our ability to analyze and understand phenomena that manifest themselves in the BGP world, *e.g.,* network outages, BGP hijacking events, and selective announcements at specific locations. In other words, to link the BGP dimension to the geographical one, we need to geo-locate BGP prefixes.

Currently, a sound approach to geo-locate BGP prefixes is lacking. The most-recent approach was developed and applied more than 14 years ago, in 2005 [4]: the authors used reverse DNS lookups of addresses in a BGP prefix to determine the set of locations that the prefix geolocated to. However, this approach did not address how prefix hierarchies can confound the geolocation of BGP prefixes, leading to potentially erroneous inferences. Further, the authors reported that only half the addresses they studied had associated reverse DNS names. In the years since 2005, IP address geolocation has improved considerably and no longer relies solely upon reverse DNS. Several commercial geolocation databases exist today, offering new tools for geolocating addresses.

In this paper, we revisit how to geolocate BGP prefixes. We show that the geo-location of prefixes is not a trivial extension of IP geo-location, *i.e.,* there is more to it than geo-locating a set of IP addresses. Unlike IP addresses, prefixes can overlap and, depending on routing dynamics and IP geo-location, their resulting geographical interpretation can vary. We formally define this problem as the *prefix geolocation ambiguity'* problem, quantify its extent, and characterise when prefixes are ambiguous.

Next, we present best-effort mappings of BGP prefixes to their locations using state-of-the-art IP address geolocation databases (NetAcuity [3] and Maxmind Geolite [13]). We study how many countries, regions, and cities prefixes geo-locate to and find, to our surprise, that 98% (GeoLite) and 99% (NetAcuity) of prefixes map to a *single* country while 60% (GeoLite) and 73% (NetAcuity) map to a *single* city. Unsurprisingly, /24s are more likely to map to a single location than less specific prefixes, and we observe a long tail of prefixes mapping to numerous locations. We manually inspect some of these outliers and find that a prefix's locations often tell an interesting tale about its use: national ISPs operate prefixes that map to a single country but numerous cities; VPN providers have long prefixes that map to numerous (sometimes spoofed!) countries; and some cloud providers and military announce short prefixes that map to only a single city. Our code and datasets are publicly available.

Our paper is organized as follows. Section II discusses the dataset of BGP prefixes and IP geolocation databases used in our analyses. In Section III, we define, quantify, and characterize the problem of prefix geolocation ambiguity. We analyze the mappings of BGP prefixes to their geolocations in Section IV. Section V proposes next steps towards a more comprehensive understanding. Section VI contrasts this study with past work that teaches us what to expect from geo-location accuracy and geo-location spoofing. Finally, Section VII concludes our work.

II. Datasets

In this section, we present the datasets we use to study the geolocation of BGP prefixes. We describe how we extract the BGP prefixes that we will geo-locate (Section II-A) and how we geo-locate IP addresses (Section II-B).

A. Extracting globally visible prefixes

To build a dataset of all network prefixes that are announced on BGP and are globally visible (*i.e.,* visible to most BGP routers on the Internet) we leverage BGP prefix reachability data from the RIPE RIS [17] and RouteViews [18] projects. Both projects operate measurement infrastructure based on

BGP collectors that peer with hundreds of operational BGP routers worldwide (BGP peers, in the following). Specifically, in our analysis we process RIB (*i.e.,* routing table) dumps [15, § 2] collected at midnight (UTC) on March 26, 2018 for all collectors of RouteViews and RIPE RIS.

We use data only from peers sharing their entire routing table with the collector—commonly called *full-feed* peers. Similarly to Orsini et al. [15, §2,5], we label peers as full-feed if they advertise more than 400,000 prefixes and partial-feed otherwise. We empirically determine this threshold by inspecting the distribution of the number of prefixes that all BGP peers see: we find a bimodal distribution, consisting of several peers with a small number of prefixes on one extreme, and peers with more than 600,000 prefixes on the other extreme; this distribution is consistent with the distribution seen by Orsini et al. [15, Fig 5(a)]. As a result, we process data from 347 out of 741 routers distributed worldwide. We then consider a prefix globally visible if at least 50% of these routers advertise it, obtaining 681,345 (75.6%) globally visible prefixes for March 26, 2018, out of 900,952 total. We also keep track of a prefix's originating autonomous system(s) as we will incorporate it in our analysis in Section III-B.

We implemented our processing code in Python, allowing us to use the Python bindings of the BGPStream library [1]. BGPStream provides a clean interface to BGP data by abstracting away the complex interaction with BGP collectors.

B. Geo-location databases

Next, we describe the geo-location databases we use to geolocate individual addresses within globally visible prefixes. We draw on two geo-location providers: NetAcuity Edge [3] (henceforth referred to as NetAcuity) and MaxMind GeoLite [13] (GeoLite). NetAcuity's commercial database has an alleged accuracy of 99.9% on a country level and 97% on a city level [3]. GeoLite in contrast is free—it is a less accurate version of the commercial GeoIP2 database that is maintained by the same company. While GeoLite was available at the time of our analysis, it has since been deprecated in favor of GeoLite2 but to the best of our knowledge, only the database format has changed. Depending on the country, GeoLite2's alleged accuracy ranges from 19% for Algeria to 100% for Singapore [12]. We decided to use both a commercial and a free database because the commercial NetAcuity allows us to investigate our problem using state of the art accuracy while GeoLite shows us what users of the free database (likely a larger user base) would experience.

We use a NetAcuity dump that was published on Mar 25, 2018. Our GeoLite dump was published on Mar 27, 2018—two days later. Recall that we extracted our globally visible prefixes on Mar 26, 2018, so our dataset and our two geo-location dumps cover a date range of only three days. This way we minimize inaccurate results caused by outdated geo-location databases. To map a prefix to its geo-location(s) we use a library that we developed ourselves. It provides a straightforward geo-location API by abstracting away the details of interacting with NetAcuity's and GeoLite's database format. For a given prefix, we find the geolocations of all of its addresses and place these locations in a set. As we will show in Section III and Section IV, this set of locations is an upper bound on the number of locations that the prefix geolocates to.

In this work, we do not study the accuracy of IP address geo-location databases, focusing instead upon using these databases to study BGP prefix geolocation, *i.e.,* the problem of assigning geo-location information to a prefix globally visible on BGP given the geo-location information of their IP addresses is known. We recognize that geo-location databases can sometimes be inaccurate and we discuss this problem in more detail in Section VI. As IP address geolocation improves, the results from our method to geolocate BGP prefixes will benefit as well.

III. Ambiguities in prefix geo-location

In this section, we show that geo-locating BGP prefixes is not a trivial extension of IP address geo-location, *i.e.,* a prefix should not necessarily be associated with the set of *all* geographical locations its respective IP addresses map to. The source of this problem is that, unlike IP addresses, announced prefixes can overlap with each other, which complicates the picture.

Consider Figure 1. A prefix a.b.c.0/23 overlaps with a separately-announced, more specific prefix a.b.d.0/24. The non-overlapping part (*i.e.,* its IP addresses) of the containing prefix maps to Belgium, whereas the overlapping addresses map to Germany. Naive geo-location of the containing prefix a.b.c.0/23 would associate it with both Belgium and Germany. However, without additional knowledge, it is unclear if the router(s) announcing the containing prefix can in fact route traffic to the overlapping addresses in Germany. For example, if prefix a.b.d.0/24 is withdrawn, can we still expect the addresses in a.b.d.0/24 to be globally reachable given that they are covered by prefix a.b.c.0/23? The answer to this question depends on the routers and ASes originating the respective prefixes, their relationship (*e.g.,* are the ASes in a directly-connected customer/provider relationship?), internal routing configuration, and physical topology. In other words, if the prefix a.b.c.0/23 (and the router announcing it) cannot provide actual connectivity for the addresses in a.b.d.0/24, then its geo-location should not include Germany. We therefore argue that if no additional information is available, the geo-location of prefix a.b.c.0/23 is *ambiguous*.

We defer the resolution of ambiguous prefixes to future work but we note that performing a rigorous mapping has practical implications when used in Internet monitoring and data analysis. As an example, consider the problem of performing live outage detection using BGP measurements: state-of-the-art systems track what prefixes are visible on BGP by a given minimum number of operational routers distributed worldwide, and use IP geo-location to perform country-level outage detection [14], [2]. CAIDA's outage detection system, IODA [2], spots an outage if the number of addresses that geo-locate to a country *and* are visible on BGP drops [2]. This

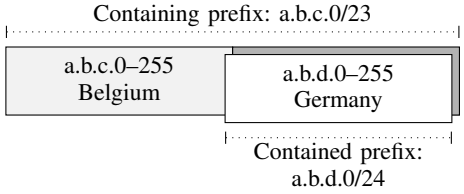


Fig. 1. A containing prefix a.b.c.0/23 whose separately announced and more specific prefix a.b.d.0/24 maps to Germany while the remainder of the containing prefix maps to Belgium.

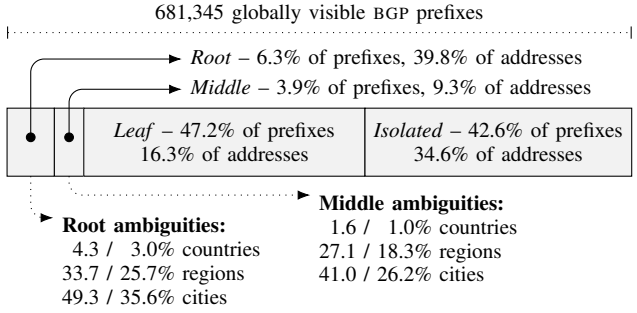


Fig. 2. Among our globally visible prefixes we distinguish between root, middle, leaf, and isolated prefixes. ↓ refers to geo-location differences to prefixes lower in the hierarchy while ↑ refers to prefixes higher in the hierarchy. For example, 49.3% (NetAcuity) and 35.6% (GeoLite) of root prefixes exhibit city ambiguities.

approach can fail when an *ambiguous* prefix is not correctly geo-located. Let us consider again the example in Figure 1 and let us assume that prefix a.b.d.0/24 (normally stable) is suddenly withdrawn. If prefix a.b.c.0/23 should theoretically be associated only with Belgium, but we instead use naive prefix geo-location (*i.e.,* mapping it to both Belgium and Germany), IODA's approach may mistakenly miss an outage in Germany. With accurate prefix geo-location available, the algorithm could take into account that once prefix a.b.d.0/24 is withdrawn, there are no prefixes geo-located to Germany keeping these addresses reachable.[1]

In the remainder of this section we first quantify what fraction of our globally reachable prefixes and address space is affected by geo-location ambiguities (Section III-A) and then seek to characterize when these ambiguities happen (Section III-B).

A. Quantifying geo-location ambiguities

How many of our 681,345 prefixes are ambiguous? To answer this question, it helps to first shed light on prefix hierarchies. To this end, we use the dataset of globally visible prefixes we assembled in Section II-A and assign each prefix to one of four mutually exclusive categories:

1) **Root** prefixes contain other (middle or leaf) prefixes but are not contained by any prefixes.
2) **Middle** prefixes both contain (middle or leaf) prefixes and are contained by (root or middle) prefixes.
3) **Leaf** prefixes are contained by (root or middle) prefixes but do not contain prefixes themselves.
4) **Isolated** prefixes are not part of the prefix hierarchy and hence do not overlap with any other prefixes.

The rectangle in Figure 2 illustrates the percentage of prefixes that fall into each category. More than half of our prefixes are part of the hierarchy, *i.e.,* they are either a root, middle, or leaf prefix. The remaining 43% of prefixes are isolated. Unsurprisingly, root prefixes constitute a small percentage of prefixes but a disproportionately large percentage of addresses. The inverse holds for leaf prefixes: they constitute almost half of all prefixes, yet only 16% of addresses.

Note that only root (6%) and middle prefixes (4%) can be ambiguous because only they contain other prefixes. But

[1]The alternative approach of simply monitoring the number of prefixes geo-located to a country [14], would similarly fail if a.b.c.0/23 was suddenly withdrawn, by mistakenly inferring an outage in Germany.

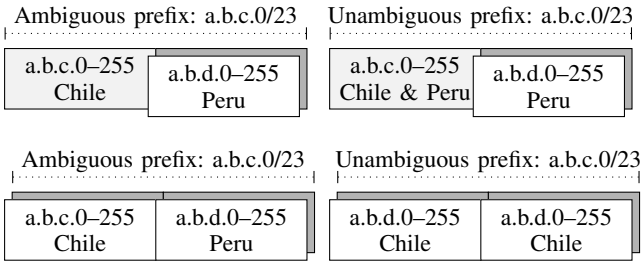


Fig. 3. The left prefixes' geo-location is ambiguous (should a.b.c.0/23 geo-locate to Peru?) while the right prefixes geo-locate unambiguously.

exactly how many prefixes among these 10% are ambiguous? To find out, we first geo-locate each prefix by mapping all IP addresses in its non-overlapping part to a set of its respective countries, regions[2], and cities. In the example of Figure 1, a.b.d.0/24 would map to {Germany} and a.b.c.0/23 would only map to {Belgium} because we discard its overlapping part—a.b.d.0/24. We then label a prefix as "ambiguous" if the locations of any of its contained prefixes (which can be a leaf or a middle prefix) is not a subset of its own locations. Algorithm 1 depicts pseudo code of the algorithm we use. Figure 3 illustrates four examples. We label the two prefixes on the left as ambiguous, because it is not clear what locations the containing prefix should geo-locate to. Should the top left prefix geo-locate to Peru? Should the bottom left prefix geo-locate to Chile? Or Peru? Or both? The two containing prefixes on the right however are unambiguous. The top right prefix geo-locates to Chile & Peru ({Peru} $\subseteq$ {Chile, Peru}) while the bottom right prefix geo-locates to Chile.

The dotted arrows branching off of the rectangle in Figure 2 point to the number of geo-location ambiguities. The percentages for the prefix types are based on our 681,345 globally visible prefixes while the percentages under the geo-location differences are based on the number of respective prefix types. The slash between percentages separates NetAcuity (on the left) from GeoLite (on the right). For example, 49.3%

[2]We use "regions" to indicate sub-national administrative divisions, such as states in the U.S., provinces in Canada, and régions in France.

Algorithm 1 Determining whether a prefix is ambiguous.

```
1: procedure IS_PREFIX_AMBIGUOUS(ctg_pfx, ctd_pfxs)
2:     ▷ Determine if the containing prefix (ctg_pfx) is ambiguous
   given the set of its contained prefixes (ctd_pfxs).
3:     ctd_loc ← GEOLOCATE(ctd_pfxs)    ▷ Determine contained
   locations.
4:     if IS_FULLY_SUBANNOUNCED(ctg_pfx) then
5:         if NUM_LOCATIONS(ctd_loc) > 1 then
6:             return True
7:         else
8:             return False
9:         end if
10:    else
11:        nonoverlapping_pfxs ← SUBTRACT(ctg_pfx, ctd_pfxs)
12:        ctg_loc ← GEOLOCATE(nonoverlapping_pfxs)    ▷
   Determine containing locations.
13:        if ctd_loc ⊂ ctg_loc then
14:            return False
15:        else
16:            return True
17:        end if
18:    end if
19: end procedure
```

TABLE I
THE PERCENTAGE OF PREFIXES (TOP ROW) AND ADDRESS SPACE OCCUPIED BY THESE PREFIXES (BOTTOM ROW) THAT ARE AMBIGUOUS WITH RESPECT TO COUNTRIES, REGIONS, OR CITIES.

	NetAcuity (%)			GeoLite (%)		
	Count.	Reg.	Cit.	Count.	Reg.	Cit.
Ambiguous prefixes	0.3	3.2	4.7	0.2	2.3	3.3
Address space	8.5	27.5	32.5	8.0	27.2	30.4

(NetAcuity) and 35.6% (GeoLite) of root prefixes have city ambiguities. We observe that root prefixes consistently exhibit more ambiguities than middle prefixes. Also, the more specific the location type, the more ambiguities we observe: country ambiguities are an order of magnitude less prevalent than region and city ambiguities.

Figure 2 shows the percentage of ambiguous prefixes with respect to the number of root and middle prefixes, but how many prefixes are ambiguous with respect to all of our 681,345 prefixes? In other words: What are the odds of a randomly-chosen prefix to be ambiguous? Table I provides the answer for prefixes in the first row, and for the corresponding IPv4 address space (compared to all $2^{32}-1$ IP addresses) in the second row. Only 0.3% (NetAcuity) and 0.2% (GeoLite) of all prefixes are ambiguous with respect to countries. City ambiguities are more likely with 4.7% (NetAcuity) and 3.3% (GeoLite). In contrast, a third of the IPv4 IP address space is ambiguous with respect to cities, which highlights the pervasiveness of ambiguous prefixes. This finding also serves as a reminder that a small number of prefixes can constitute a significant fraction of the address space.

B. Understanding geo-location ambiguities

Here, we take a first look into characterizing when prefixes are ambiguous and investigate the potential for using the resulting inferences to resolve geo-location ambiguities. Resolving a containing prefix's ambiguous geo-location requires determining whether the containing prefix should also be assigned the location(s) of addresses from its contained prefixes. For example, we would like to determine if the containing prefix in Figure 1 should map to both Belgium and Germany, or to Belgium alone. We present a preliminary exploration of geo-location ambiguities and present potential explanations and predictors that can help resolve ambiguities.

Our first intuition was that ambiguous prefixes may be announced by an origin AS that is different from its contained prefix(es). If an ambiguous containing prefix and its contained prefix are announced by separate ASes, we could consider the prefixes different for all practical purposes, which would resolve the ambiguity.

We iterated over all 472,174 pairs of containing[3] and contained prefixes and determined for each pair if the containing prefix is ambiguous (with respect to countries, regions, and cities) compared to its contained prefix, and if there is an origin AS difference between the two. Though only 69,264 (10%) of the 681K globally visible prefixes are containing prefixes, these containing prefixes form a large number of prefix *pairs* (472,174) with their contained prefixes. Table II shows the results of this analysis. The four percentages of each of our three location types add up to 100%. Each location type has two columns, indicating the percentage of identical geo-locations ($=$) and differing geo-locations ($\neq$). The table's rows show the percentage of identical and differing origin ASes. For example, related to countries, 20.6% of prefix pairs geo-locate to identical countries but have different origin ASes. The table shows that origin AS differences are generally less likely than identical origin ASes *except* when there are country ambiguities. We highlight this result further in Table III, which shows an alternate representation of the data from Table II. Table III shows the conditional probabilities that the origin AS differs given that there is a difference in the location of the containing and contained prefix pairs. Interestingly, we observe that an origin AS difference occurs for two-thirds of the containing and contained prefix pairs which have a country ambiguity. For region and city ambiguities, such origin AS differences are less likely.

Next, we hypothesized that shorter containing prefixes are more likely to be ambiguous, since they can contain many longer prefixes which may geolocate to a different location. We confirm this hypothesis by examining the relationship between the length of the containing prefix and the likelihood that the prefix is ambiguous at the country and city levels in Figure 4. Shorter prefixes are indeed more likely to be ambiguous.

We then investigated which autonomous systems are particularly likely to have prefixes that exhibit geo-location ambiguities. Figure 2 shows that our dataset has 69,264 containing prefixes; of these, 2,269 exhibit a country-level ambiguity with at least one of their contained prefixes. The ten ASes with the most containing prefixes that have country-level ambiguities

[3]The set of root and middle prefixes together constitutes containing prefixes.

TABLE II
THE RELATIONSHIPS BETWEEN THE GEO-LOCATION AND ORIGIN AS(ES) OF A CONTAINING AND ITS CONTAINED PREFIXES.

	NetAcuity (%)						GeoLite (%)					
	Country		Region		City		Country		Region		City	
	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$
$\neq$ origin AS	7.4	20.6	18.6	9.4	22.4	5.6	5.9	22.1	18.0	10.0	20.4	7.6
$=$ origin AS	3.7	68.3	39.5	32.5	49.0	23.0	3.0	69.0	33.9	38.1	41.2	30.8

TABLE III
THE CONDITIONAL PROBABILITIES OF AN ORIGIN AS DIFFERENCE KNOWING THAT THERE IS A GEO-LOCATION AMBIGUITY.

Conditional probability	NetAcuity	GeoLite
P(origin-as-diff | country-diff)	0.66	0.66
P(origin-as-diff | region-diff)	0.32	0.35
P(origin-as-diff | city-diff)	0.31	0.33

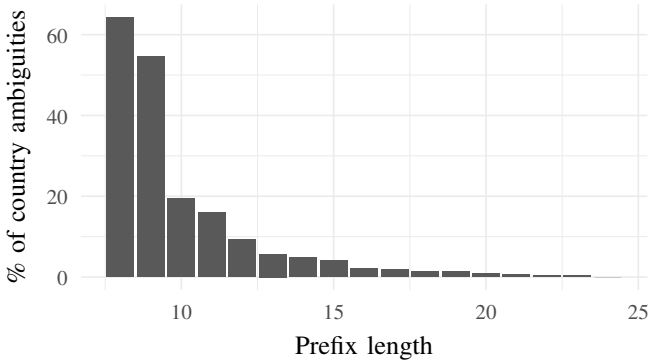
include large transit providers (Level 3, Cogent, Open Transit), large residential ISPs (Centurylink, Verizon), software companies (Apple, Accenture), and government agencies (U.S. DoD). These ten ASes account for 283 (12.4%) of the 2,269 prefixes.

For most of the containing prefixes from the above ASes, their contained prefixes were announced by either the same AS or frequently, by a sibling AS. In these cases, it is plausible that the containing prefix can provide connectivity to the contained prefix's addresses if the latter is withdrawn; thus, the containing prefix should likely inherit the location(s) of the addresses from its contained prefixes. However, for some of the above ASes, we observe that the containing and their respective contained prefixes do not appear related to each other, *i.e.,* the prefix pairs belong to different ASes and these ASes are neither siblings, nor are they in a customer-provider relationship. For example, of 30 containing prefixes belonging to Level 3 that were ambiguous at the country level, 25 have contained prefixes announced by ASes that appear unrelated to Level 3[4]. For such containing prefixes, we may be able to resolve the geo-location ambiguity by only considering the geo-locations of addresses that the prefixes can feasibly provide connectivity to (*i.e.,* if the contained prefix belongs to a sibling or customer AS).

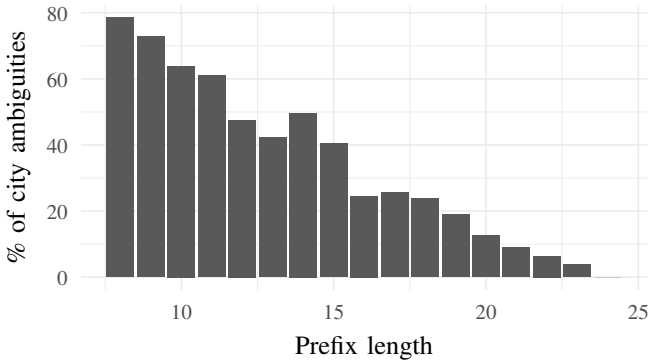
Our preliminary results suggest that it may be possible to resolve ambiguities in containing prefixes' geo-locations by examining various features associated with the containing and contained prefixes, such as their origin ASes and prefix lengths. In future work we will examine whether additional features can help explain geo-location ambiguities, *e.g.,* the type of ASes that the containing and contained prefixes belong to (transit, stub, multi-homed), the extent to which the containing prefix is subannounced etc.

In this paper, we handle ambiguous prefixes by reporting the *upper bounds* of the number of locations that these prefixes geo-locate to. In Section IV, we geo-locate prefixes by having them map to *all* contained locations—even if the prefix is ambiguous. Our results therefore provide an upper bound on the number of locations that a prefix geo-locates to.

[4]It is possible that Level 3 shares a relationship with these ASes that we are unaware of.



(a) The % of country ambiguities by containing prefix length.



(b) The % of city ambiguities by containing prefix length.

Fig. 4. The percent of country (top) and city (bottom) ambiguities by containing prefix length.

IV. GEO-LOCATING PREFIXES

Next, we arrive at another key question of our work: *What places do prefixes geo-locate to?* Section IV-A first takes a look at how many locations (*i.e.,* countries, regions, and cities) prefixes geo-locate to and Section IV-B then studies interesting outliers that we discovered.

A. How many locations do prefixes geo-locate to?

We begin by iterating over the globally visible prefixes we extracted in Section II-A and mapping each IPv4 address

to its location, using both NetAcuity and GeoLite. Both geo-location databases provide information about an address' country, region, and city. For each prefix, we count the number of countries, regions, and cities it maps to. Note that we do not distinguish between ambiguous and unambiguous prefixes for this analysis; thus, the results we report here describing the number of locations that a prefix maps to are upper bounds.

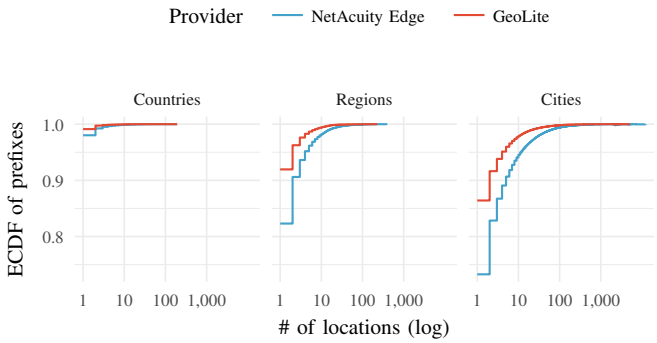
Figure 5 illustrates the results. In the top three diagrams, the x-axis shows the number of locations (in $\log_{10}$) while the y-axis represents the empirical CDF of prefixes that map to the given number of locations. Interestingly, it is unlikely for a prefix to map to more than one geographical location—be it a country, region, or even a city. However, this depends on a prefix's size, as we will show later. Most prefixes map to a single location: 98% (NetAcuity) and 99% (GeoLite) geo-locate to a single country, 80/67% to a single region, and 73/60% to a single city. The bottom three diagrams in Figure 5 illustrate the number of IP addresses that are covered by the prefixes constituting the top three diagrams. The number of addresses refers to all addresses covered by the respective prefixes. We count addresses that are covered by multiple prefixes only once, *i.e.,* the prefixes a.b.c.0/23 and a.b.c.0/24 cover 512 addresses. All sub-plots in Figure 5 have a long tail: 0.3% (NetAcuity) and 0.1% (GeoLite) of prefixes map to more than five countries while 5.4/2.1% map to more than ten cities. Section IV-B will discuss the most interesting outliers we discovered. These numbers show that BGP prefixes can vary in the number of locations they geolocate to, but many map to a single location. This result suggests that we can potentially use BGP data to detect and analyze phenomena, such as connectivity outages, even at fine geographic granularity.

Figure 5 treats all prefixes equally and does not differentiate by their length. Our intuition however tells us that prefix length matters—a /8 is likely to map to more locations than a /24. We therefore repeat our analysis for all prefix lengths in our dataset, ranging from /8 (the shortest) to /24 (the longest). Figure 6 illustrates the number of geographical locations that prefixes of a given length map to. Indeed, we find that it confirms our intuition: more specific prefixes geo-locate to fewer locations. Also, prefixes map to fewer countries than regions, and fewer regions than cities. Still, even /24s occasionally map to more than one location—0.5% map to more than one country and 5.8% map to more than one city.

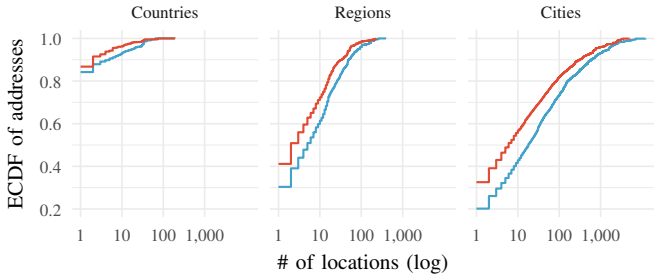
Recall that this analysis did not take into account that some prefixes are ambiguous. Thus, the number of locations that prefixes geolocate to in Figure 5 and Figure 6 are an upper bound.

B. Geo-location outliers

Figure 6 tells us how many locations a prefix of a given size tends to map to but not all prefixes adhere to these expectations. We will now take a closer look at prefixes that geo-locate to an unexpected number of locations. We find that our data tells interesting tales about the purpose of prefixes and, more broadly, provides a new lens to look at Internet data. We do not simply rank all prefixes by the number of



(a) The empirical CDF of prefixes over the number of locations.



(b) The empirical CDF of prefixes' addresses over the number of locations.

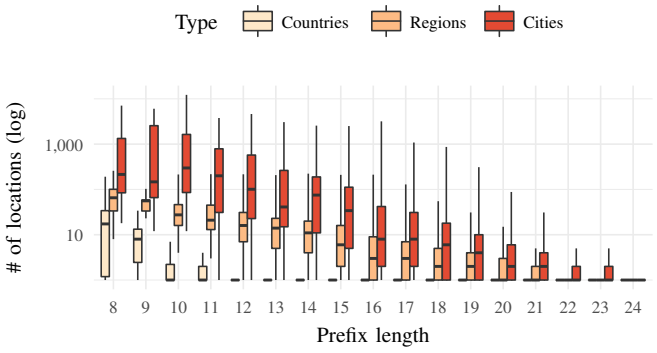
Fig. 5. The y-axis represents the empirical CDF of prefixes (top) and the prefixes' covered addresses (bottom) that geo-locate to a given number of countries, regions, and cities (x-axis).

locations they map to because short prefixes such as a /8 tend to map to more locations than long prefixes. We therefore use a *location diversity metric* to explore outliers, which we define as:

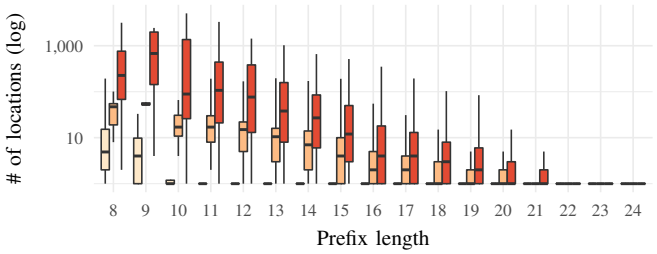
$$\mathscr{D}_{\text{location}} = \frac{\text{\# of locations}}{\text{\# of reachable addresses}}$$

A /24 prefix whose addresses geo-locate to 254 different countries would have a diversity metric of $254/254 = 1$. If it only geo-locates to a single country, the metric would be $1/254 = 0.004$. Note that our diversity metric exhibits bias: theoretically, a /24 can have maximum diversity for its 254 addresses but a /8 prefix's diversity metric can never equal 1 because there are no 16 million countries. Regardless, we find that our metric is useful for exploratory analysis, and indeed, we discovered three noteworthy outlier classes: (*i*) prefixes that geo-locate to *many* countries but *few* cities; (*ii*) prefixes that geo-locate to *many* cities in a *single* country; and (*iii*) short prefixes that geo-locate to a *single* city.

1) Prefixes in many countries but few cities: We discovered many long prefixes that geo-locate to numerous countries while being present in only one city per country. Upon inspecting the originating autonomous system and the whois records, we discovered that these prefixes, shown in Table IV, are owned by two VPN providers: "HideMyAss!" (a subsidiary of AVAST) and "IAPS Security Services" whose prefix is announced by NFORCE. Users often use VPNs to



(a) NetAcuity



(b) GeoLite

Fig. 6. The number of geographical locations that prefixes with a given length geo-locate to. We distinguish between countries (lightest), regions (darker), and cities (darkest). The box plots do not contain outliers beyond the whiskers.

TABLE IV
PREFIXES THAT GEO-LOCATE TO MANY COUNTRIES BUT ONLY A FEW CITIES. THE PREFIXES ARE OWNED BY TWO VPN PROVIDERS.

Prefix	# countries	# cities	# city errors	Owner
5.62.60.0/23	127	100	27	HideMyAss!
5.62.62.0/23	125	100	26	HideMyAss!
5.62.56.0/23	68	72	23	HideMyAss!
5.62.58.0/23	62	67	20	HideMyAss!
46.36.203.0/24	28	23	33	IAPS
46.36.201.0/24	24	46	18	IAPS
46.36.200.0/24	23	31	11	IAPS

circumvent location-based content restrictions, which incentivizes VPN providers to maximize their country footprint. The prefix 5.62.60.0/23 maps to 127 different countries—including Christmas Island, North Korea, and even Vatican City. Our skepticism about these locations was reinforced by the work of Weinberg et al. whose IMC'18 paper showed that some VPN providers "lie" about their proxies' geo-location [20]. Earlier, in 2014, a blogger also expressed doubts about the alleged locations of IAPS's proxies [8]. All prefixes in Table IV contain several IP addresses that NetAcuity could not geo-locate—a particularly rare error considering NetAcuity's high coverage.

Apart from the VPN providers in Table IV, the prefixes with the largest number of countries are dominated by the French cloud hosting provider OVH and the Swedish hosting provider MissDomain.

2) Prefixes in many cities but one country: Figure 6 shows that shorter prefixes tend to geo-locate to more locations. Our data exhibits many (both short and long) prefixes that geo-locate to one country but hundreds, or even thousands of cities. Unsurprisingly, such prefixes mostly belonged to large national ISPs such as the German Deutsche Telekom, the American Comcast, and the French Orange. We also find long prefixes that map to a surprisingly large number of cities. The French telecommunication provider Orange announces several /24 and /23 that map to hundreds of French cities. For example, 62.160.124.0/24 geo-locates to 177 cities.

3) Short prefixes in a single city: Finally, we stumbled upon prefixes as short as a /8 that geo-locate to *a single city*. Our data shows that many prefixes—one as short as /8—geo-locate to a single city. Upon manual inspection we find that these outliers mostly belong to telecommunication providers, cloud providers, and military: (*i*) The Kenyan provider Safaricom operates two /12 prefixes that geo-locate to Nairobi, (*ii*) Korean Telecom operates several /12 prefixes that geo-locate to Seoul, (*iii*) the Chinese TieTong Corporation operates a /11 and /12 that map to Beijing, and (*iv*) the German Deutsche Telekom announces a /11 that maps to Bonn. Amazon announces a /12 that maps only to Ashburn in the U.S. (Amazon operates a data center in the city); Samsung announces a /12 that maps to Seoul; and Cloudflare announces a /12 that maps to San Francisco; the Dutch research and education service provider SURFnet announces a /12 that maps to the Dutch city Utrecht. Finally and most strikingly, the U.S. Army Intelligence and Security Command announces 55.0.0.0/8, which maps to Fort Huachuca, a U.S. Army installation.

V. DISCUSSION

A *longitudinal analysis* would provide further insight into how the geographical properties of prefixes change over time. We only considered a single snapshot in time but GeoLite (but not NetAcuity) maintains an archive of past database releases, making it straightforward to rerun our experiments over past data. Finally, our work is limited to IPv4 prefixes. Future work should study the geo-location of IPv6 prefixes, and highlight how these results compare to our IPv4-based results.

Section III-B lamented the lack of *ground truth* that has the correct mapping of a prefix to its geo-locations, regardless of potential ambiguities. One could create ground truth by repeatedly running traceroutes to a contained prefix that is causing ambiguity for its containing prefix. Once the contained prefix is withdrawn, the behavior of traceroutes could turn ambiguity into certainty: if the traceroutes keep reaching their destinations we may conclude that the containing prefix also geo-locates to the overlapping locations. If however the traceroutes do not reach their destinations, we may not be able to draw conclusions, since the withdrawal might be related to events of different nature, including failures in the final destination network. Large-scale, longitudinal traceroutes coupled with BGP data may provide the reference information we need to resolve these ambiguities.

VI. Related work

Most similar to this paper is Freedman et al.'s [4] work from IMC'05, in which they studied the geographical locality of prefixes and their effect on route aggregation, routing table size, and routing policies. In contrast to our use of geo-location databases, the authors relied on reverse DNS to map an IP address to its location. Freedman et al.'s work counts more than ten years of age—routing tables have since grown and changed substantially—and did not consider prefix hierarchies.

In this paper, we did not evaluate the accuracy of the geo-location databases we used; however, this topic has received attention from the research community [16], [19], [5], [9]. Geo-location appears to be highly accurate at the country level. In 2011, Poese et al. [16] reported a 96–98% accuracy depending on the database. Similarly, Shavitt et al. [19] reported at least 80% for seven databases, also in 2011. Finally, Gharaibeh et al. [5] showed at IMC'17 that their tested databases are at least 77.5% accurate when geo-locating Internet routers (NetAcuity had 89.4% accuracy at the country-level). Note however that these works used very different datasets for ground truth, so the accuracy numbers cannot be compared. City-level accuracy has lagged behind country-level accuracy, although recent work suggests that accuracy is improving at the city-level too. In 2011, Poese et al. [16, § 3] measured that "less than 20% of the exact matches for Maxmind and InfoDB are within a few tens of Km from the ground truth." In the same year, Shavitt et al. [19, § IV.B] measured city-level accuracy ranging from as low as 0.8% (IPligence) to 79.1% (NetAcuity). In 2017, however, Gharaibeh et al. [5, § 5.2.1] showed that the NetAcuity geolocation database was successfully able to geolocate router addresses at the city-level with an accuracy of 74.2%. Considering that Gharaibeh et al. focused on geolocating router addresses, and that geolocation databases are especially bad at it (since their primary focus is on the edge [10]), we expect the geolocation accuracy of non-router addresses to be better.

There are incentives to spoof one's geo-location, *e.g.,* to bypass region blocks for online video. Not only is it possible to interfere with active geo-location methods [6] but Weinberg et al. [20] recently showed that it is happening in the wild: several VPN providers spoof geo-location to make their customers believe that they operate proxies in countries they are not. Our results from Section IV-B offer additional evidence that some VPN providers may be spoofing their geo-location.

VII. Conclusion

In this work we took a first look at geo-locating BGP prefixes. This problem differs from (but builds on) geo-locating IP addresses because of the complexity that BGP introduces. We showed that simply mapping each of a prefix's IP addresses to its geographical location may be misleading because separately announced, more specific prefixes, can map to different locations. We quantified how often this happens for both the commercial NetAcuity Edge and the free GeoLite database and took preliminary steps in the investigation of how to solve such ambiguous cases. We also found that most prefixes map to a single geographical location—for NetAcuity 98% map to a single country and 73% to a single city. This is an interesting result, because it suggests that BGP data analysis can reveal phenomena (*e.g.,* connectivity outages) even at a relatively fine geographic granularity. It is likely that such potential will further increase in the future, given the increasing fragmentation of prefixes in the global routing table and the fact that (as we found) longer prefixes are more likely to map to a single location than short ones.

VIII. Acknowledgments

This research is supported by the U.S. Department of Homeland Security S&T Directorate via contract number 70RSAT18CB0000015, by the Air Force Research Laboratory under agreement number FA8750-18-2-0049 and by NSF grants CNS-1705024 (MapKIT), OAC-1724853 (PANDA), OAC-1848641 (ARTEMIS) and CNS-1423659 (HIJACKS).

References

[1] CAIDA. BGPStream. [Online]. Available: https://bgpstream.caida.org
[2] CAIDA. IODA – Internet Outage Detection and Analysis. [Online]. Available: https://ioda.caida.org
[3] Digital Element. NetAcuity edge premium. [Online]. Available: https://www.digitalelement.com/solutions/netacuity-edge-premium/
[4] M. J. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *IMC*. USENIX, 2005.
[5] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A look at router geolocation in public and commercial databases," in *IMC*. ACM, 2017.
[6] P. Gill, Y. Ganjali, B. Wong, and D. Lie, "Dude, where's that IP? circumventing measurement-based IP geolocation," in *USENIX Security*. USENIX, 2010.
[7] B. Gueye, S. Uhlig, and S. Fdida, "Investigating the imprecision of IP block-based geolocation," in *PAM*. Springer, 2007.
[8] hacker10. Review scam vpn provider iaps intl-alliance. [Online]. Available: https://hacker10.com/other-computing/review-scam-vpn-provider-iaps-intl-alliance/
[9] B. Huffaker, M. Fomenkov, and kc claffy. (2011) Geocompare: a comparison of public and commercial geolocation databases. [Online]. Available: https://www.caida.org/publications/papers/2011/geocompare-tr/geocompare-tr.pdf
[10] ——. (2014) DRoP: DNS-based router positioning.
[11] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *IMC*. ACM, 2006.
[12] MaxMind. GeoIP2 city accuracy. [Online]. Available: https://www.maxmind.com/en/geoip2-city-accuracy-comparison
[13] ——. GeoLite legacy downloadable databases. [Online]. Available: https://dev.maxmind.com/geoip/legacy/geolite/
[14] Oracle. Internet Intelligence Map. [Online]. Available: https://internetintel.oracle.com/
[15] C. Orsini, A. King, D. Giordano, V. Giotsas, and A. Dainotti, "BGPStream: A software framework for live and historical BGP data analysis," in *IMC*. ACM, 2016.
[16] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP geolocation databases: Unreliable?" *Computer Communication Review*, vol. 41, no. 2, 2011.
[17] RIPE NCC. Routing information service (RIS). [Online]. Available: https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris/routing-information-service-ris
[18] Routeviews. University of oregon route views project. [Online]. Available: http://www.routeviews.org
[19] Y. Shavitt and N. Zilberman, "A geolocation databases study," *Selected Areas in Communications*, vol. 29, no. 10, 2011.
[20] Z. Weinberg, S. Cho, N. Christin, V. Sekar, and P. Gill, "How to catch when proxies lie: Verifying the physical locations of network proxies with active geolocation," in *IMC*. ACM, 2018.