How Accurate is IP Geolocation Based on IP Allocation Data?

Sebastian Zander
Centre for Advanced Internet Architectures, Technical Report 120524A*
Swinburne University of Technology
Melbourne, Australia
szander@swin.edu.au

***Abstract*—The Regional Internet Registries (RIRs) publish information about all allocated IPv4 and IPv6 address blocks including country codes for each block. This information can be used for IP address geolocation. However, since large IP address blocks assigned to one country may belong to large international organisations spread over multiple countries the accuracy is questionable. With data from May 2012 we compare the IP to country mapping of the RIR data for the whole allocated IPv4 and IPv6 space against MaxMind's GeoLite country geolocation database, which has a claimed accuracy of 99.5% [1]. For IPv4 there is a difference for 5% of the address space. This means GeoLite is presumably more accurate, but on the other hand for 95% of the IPv4 address space the mapping is identical. For IPv6 there is almost no difference between the RIR data and GeoLite.**

***Index Terms*—IP Geolocation, RIR Allocated IPs**

I. Introduction

Five Regional Internet Registries (RIRs) manage the distribution of Internet number resources including IP addresses and Autonomous System Numbers (ASNs) [2]. Each RIR consists of the Internet community in one region: AfriNIC (Africa), APNIC (South East Asia, Pacific), ARIN (North America), LACNIC (Middle/South America), and RIPE NCC (Europe, Northern Asia). The Number Resource Organization (NRO) is a coordinating body for the five RIRs [2].

The RIRs publish information about all allocated/assigned IPv4 and IPv6 address blocks, which we will refer to as *delegated* data [2]. The delegated data contains ISO 3166-1 2-digit country codes for each block specifying the country of the allocation that can be used for IP address geolocation. IP address geolocation maps IP addresses to their geographic locations, usually identified by latitude/longitude coordinates. IP address geolocation is useful for a number of applications, such as customised content, targeted online advertisements, fraud detection, and web server statistics.

The delegated data only allows IP address geolocation with country-level granularity, where the coordinates are basically the geographic "centres" of the countries. A more fine-grained mapping of IP addresses to geographical coordinates of individual cities is not possible, but often it is also not needed. However, IP geolocation based on the delegated data is likely inaccurate, since large IP address blocks assigned to one country may belong to large international organisations that are spread over multiple countries. In this report we investigate how accurate the delegated data is when compared to a presumably accurate dedicated IP geolocation database of a commercial provider.

As reference data, resembling the "ground truth", we chose MaxMind's freely available GeoLite country databases for IPv4 and IPv6 [1]. According to MaxMind their IPv4 database is very accurate on country level. MaxMind claims the accuracy of GeoLite country for IPv4 is 99.5% [1] – only slightly lower than the claimed 99.8% accuracy of MaxMind's non-free GeoIP country database for IPv4 that is used by a number of large companies [3]. MaxMind's IPv6 database is still under development and its accuracy is unpublished.[1]

Given the size of the allocated IP number space, especially for IPv6, we cannot compare the delegated and GeoLite data by querying individual IP addresses. Instead we developed an algorithm that compares the whole space on a block by block basis, where in the delegated data a block is a block of allocated IP addresses and in the GeoLite data a block is a block of consecutive IP addresses with the same country code.

*This tech report was revised on the 30th of November 2012. We replaced the incorrect mention of "GeoIP" with the correct "GeoLite" in a number of places.

[1]Private email exchange with MaxMind's support staff.

The rest of the report is organised as follows. Section II describes our methodology and Section III presents the results. Section IV concludes and outlines future work.

II. Methodology

Our study is based on the delegated data from the 20th of May 2012 (text files plus database), which we compare against the GeoLite databases updated on the 1st of May 2012. Since the GeoLite database is relatively recent compared to the delegated data we expect that there are not many recently allocated blocks missing.

Our software parses the GeoLite ASCII database and the delegated data. For both it creates a list of blocks, where each block is a tuple of <start IP address, end IP address, country code>. The lists are ordered by the start-addresses. The first and last IP addresses are stored as integer numbers. The GeoLite database already contains the first and last addresses as integer numbers, and for the delegated data we convert the ASCII IP addresses to integer numbers using the same method as used by GeoLite [1].[2]

Our algorithm then compares the lists of delegated blocks D and GeoLite blocks G step-by-step starting at the beginning. In each step the algorithm compares the current block d_i of D and g_j of G, and depending on the start- and end-addresses advances the current block in one or both lists. Thus the complexity of the algorithm is $\leq O(\|D\| + \|G\|)$. Algorithm 1 shows the pseudo-code for the comparison algorithm.

The algorithm classifies each comparison into one of a number classes depending on the overlapping pattern of the blocks (lines 4–29). For example, d_i and g_j may overlap exactly (identical start- and end-address), d_i may be completely contained in g_j (or vice versa), or d_i and g_j may partially overlap (e.g. d_i has higher start- and end-address than g_j or vice versa). Two special cases are a block in D that is (partially) not covered by any block in G (referred to as *missing*), and a (partial) block in G that is actually not in D (referred to as *unallocated*).

The algorithm counts the number of allocated IP addresses (line 31) and identifies IP addresses with non-matching and matching countries. In each comparison step the number of IP addresses is the size of the intersection of d_i and g_j. If country codes are different this set of IP addresses does not match; otherwise it does match (lines 33–42). Figure 1 shows an example

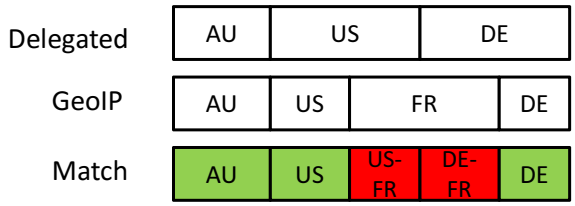

Figure 1. Delegated vs GeoLite address block comparison example

block sequence and the counted non-matching (red) and matching addresses (green).[3]

The algorithm also logs the number of IP addresses for all combinations of non-matching countries between D and G (lines 39), which is later used to compute the percentage of non-matching addresses divided by the total non-matching addresses for each country mismatch combination. Furthermore, the algorithm counts all allocated blocks in D. For each block in D (assigned to some country) where at least one IP address is mapped to a different country in G it counts the block as non-matching; otherwise it counts the block as matching (lines 44–51).[4]

Finally, the algorithm advances in the list with the block with smaller end-address or advances in both lists if d_i and g_j have identical end-addresses (lines 53–59).

III. Results

Table I shows the results of the comparison for the IPv4 address space. There is a significant difference both in terms of allocated blocks (9.7%) and IPs (5.0%). If one trusts MaxMind's claimed high accuracy, this means that GeoLite provides a more accurate mapping for this differing portion of IPv4 space. A very small percentage (0.2%) of IPs allocated is not in the GeoLite database, presumably due to GeoLite's update lag. Interestingly, we found a small percentage of IPs (0.9%) that are covered by GeoLite, but are not allocated according to the delegated data. The reasons for this are unclear.

Almost 91% of differences between the delegated data and GeoLite in terms of number of IP addresses are caused by the following mis-mappings:

- Addresses assigned to Europe (EU) by the delegated data, but mapped to (mostly) European countries in GeoLite (75.5%);

[2]For IPv6 this requires the use of 128 bit integers, e.g. implemented by Perl's BigInt package.

[3]Missing IP addresses in G are counted as non-matching whereas unallocated addresses present in G are counted separately.

[4]For the sake of brevity we omitted the code for counting missing and unallocated addresses from Algorithm 1.

Table I
COMPARISON OF COUNTRY CODES FOR DELEGATED DATA AND GEOLITE LITE FOR IPV4

	Matching	Non-matching	Missing (GeoLite)
Allocated Blocks	125301 (90.0%)	12120 (9.7%)	431 (0.3%)
Allocated Addresses	3.430G (94.9%)	0.171G (5.0%)	5.373M (0.2%)

- One entire /8 prefix assigned to a large international company mapped to France by the delegated data, but mapped to EU in GeoLite (9.8%);
- Addresses mapped to the USA by the delegated data, but mapped to different other countries by GeoLite (5.4%).

The remaining 9% are many different small mis-mappings, often to some degree involving neighbouring countries e.g. Germany-Netherlands, Portugal-Spain etc.

Table II shows the results of the comparison for the IPv6 space (numbers of allocated addresses omitted due to the very large numbers). There is only a very small difference for blocks (0.1%) and IPs (0.007%), however the percentage of allocated space not covered by GeoLite is larger than in the case of IPv4 (0.4% of the address space). The GeoLite database for IPv6 is lagging behind further, presumably due to higher allocation activity for IPv6 or slower updating of GeoLite. Again, a small percentage (0.4%) of IP addresses mapped in GeoLite are not allocated according to the delegated data.

The main difference in terms of the number of IP addresses is caused by the delegated data using the obsolete country code "Netherland Antilles" for an allocation attributed to Curacao in GeoLite (90.9%).

Assuming that GeoLite is as accurate as claimed, we can conclude that for IPv4 GeoLite is more accurate than the delegated data. However, for 95% of the address space the delegated data is equal to GeoLite, which still may be an acceptable accuracy for some applications. However, note that in any real applications the accuracy depends on the distribution of observed IPs. For example, if the distribution of observed IPs is skewed towards address blocks where GeoLite is presumably more accurate, using the delegated data will result in much lower accuracy.

For IPv6 there is not much difference between the delegated and GeoLite data. Since the delegated data is more up to date with new allocations[5], it may be better to use the delegated data. On the other hand if the time frame between allocation and actual use of address blocks is more than a few weeks, using the delegated data would not be much of an advantage.

We note that a small part of space mapped in GeoLite is not allocated according to the delegated data. This is no problem for most applications where only observed/used IP addresses are queried. However, this may be a problem in other cases, for example if one queries random addresses or all addresses.

Since our algorithm in each step logs how blocks in D and G overlap, we also briefly looked at the patterns. There are two main patterns: GeoLite mainly differentiates allocated blocks into smaller blocks with different country codes (increasing the accuracy), but it also aggregates some consecutive allocated blocks with same country code (minimising the database size).

IV. CONCLUSIONS AND FUTURE WORK

In this report we investigated how accurate the IP allocation data of the Regional Internet Registries (RIRs) is for IP address geolocation with country granularity. We compared the RIR IP address allocation data against MaxMind's GeoLite country geolocation database, which has a claimed accuracy of 99.5% [1]. We compared the mapping of IPs to countries for both datasets for IPv4 and IPv6 in May 2012.

For IPv4 there was a difference in the country mapping for about 5% of the IP address space meaning GeoLite was presumably more accurate. On the other hand the difference was not very large since for 95% of the IPv4 address space the mapping was identical. With a "uniform" distribution of observed IP addresses the 95% accuracy of the delegated data may well be sufficient. If observed IP addresses are skewed towards space where the country codes of the delegated data and GeoLite data match, the accuracy could be even higher. However, if the observed IP addresses are skewed towards space where the delegated data is not accurate, using GeoLite would result in higher accuracy.

For IPv6 the delegated data and GeoLite country database are almost identical, so the delegated data was as accurate as GeoLite (note that MaxMind's IPv6 database is still under development). The delegated data is updated more frequently than the GeoLite database,

[5]The delegated data is updated every day, but the GeoLite database is updated only once per month.

Table II
COMPARISON OF COUNTRY CODES FOR DELEGATED DATA AND GEOLITE FOR IPV6

	Matching	**Non-matching**	**Missing (GeoLite)**
Allocated Blocks	11004 (98.4%)	14 (0.1%)	158 (1.4%)
Allocated Addresses	99.6%	0.007%	0.4%

e.g. in our comparison GeoLite did not cover about 0.4% of the allocated IPv6 space. However, it is unclear if that would make a significant difference in practice, since some time passes between the allocation of address blocks and their use.

In future work we plan to analyse the trends of the country-mapping differences over time, especially changes for the IPv6 space facilitated by the uptake of IPv6 and hence the increasing relevance of IPv6 geolocation.

ACKNOWLEDGEMENTS

This research was supported under Australian Research Council's Linkage Projects funding scheme (project LP110100240) in conjunction with APNIC Pty Ltd.

REFERENCES

[1] "MaxMind GeoLite Country Database." http://www.maxmind.com/app/geolitecountry.
[2] "Number Resource Organization (NRO) web site." http://www.nro.net.
[3] "MaxMind GeoIP Clients." http://www.maxmind.com/app/geoipclients.

Algorithm 1 Address block country code comparison algorithm

```
d = get_next(D)
g = get_next(G)
while (d != END && g != END) {
        if (d == END)
                type = unalloc
        else if (g == END)
                type = notcov
        else if (d.start == g.start && d.end == g.end)
                type = full
        else if (d.start == g.start && d.end > g.end)
                type = part
        else if (d.start < g.start && d.end > g.end)
                type = gcont
        else if (d.start < g.start && d.end == g.end)
                type = gcontend
        else if (d.start == g.start && d.end < g.end)
                type = over
        else if (d.start > g.start && d.end < g.end)
                type = dcont
        else if (d.start > g.start && d.end == g.end)
                type = dcontend
        else if (d.end < g.start)
                type = notcov
        else if (d.start > g.end)
                type = unalloc
        else if (d.start > g.start && d.end > g.end)
                type = overlap
        else if (d.start < g.start && d.end < g.end)
                type = overlap2

        bsize = ips_intersect(type, d, g)

        if (d.country == g.country)
                ips_match += bsize
        else {
                if (type != unalloc)
                        block_nomatch = true
                ips_nomatch += bsize
                log_nomatch(d.country, g.country, bsize)
        }

        ips_total += bsize

        if (type in (full, over, dcont, dcontend, gcontend, overlap2)) {
                if (block_nomatch)
                        blocks_nomatch++
                else
                        blocks_match++
                block_nomatch = false
                blocks_total++
        }

        if (d.end < g.end)
                d = get_next(D)
        else if (d.end > g.end)
                g = get_next(G)
        else
                d = get_next(D)
                g = get_next(G)
}
```