Comparison of GeoIP Databases for Tor

Karsten Loesing
karsten@torproject.org

Tor Tech Report 2009-10-001
October 23, 2009

Abstract

Tor uses a GeoIP database to resolve client IP addresses to country codes to get some basic statistics on connecting clients per country. Two recent events indicate that the GeoIP database that Tor ships is less accurate than expected: first, the update from the June 2009 to the September 2009 database removed almost all US IP addresses probably because of a provider-side database problem; second, in the aftermath of Tor being blocked in China at the end of September 2009, not only Chinese bridge usage increased, but also usage from Japan and Australia, which is most likely a result of Chinese IP addresses falsely resolving to those countries. This report compares various, preferably free GeoIP databases for their accuracy in mapping IP addresses to country codes. In particular, accuracy is evaluated for countries that potentially censor the Internet.

1 Motivation

Tor requires a GeoIP database that is as accurate as possible for resolving IP addresses of small countries like Iran or Tunisia. Two recent events indicate that the accuracy of the currently shipped database is not sufficient.

Unreliable database updates The GeoIP database in Tor is updated every few months, so as to reflect changes. However, the update from the June 2009 to September 2009 version introduced major changes to the database, removing almost all US IP addresses. In particular, the diff between old and new versions has 19833 deletions and only 10470 insertions.[1] This likely corrupt database update indicates that the database provider is not as reliable as expected.

False classification of Chinese IP addresses Starting on September 25, 2009, the number of bridge users coming from China increased significantly as a response to the blocking of Tor relays in China. But at the same time, statistics show a significant increase of Australian and Japanese bridge usage for no good reason. The most likely explanation is that the GeoIP database falsely classifies Chinese IP addresses as belonging to either Australia or Japan. It is impossible to say whether the reason is that GeoIP databases on the bridges doing the resolution are outdated, or if the inaccuracy still persists in more recent versions.

[1] http://archives.seul.org/or/cvs/Sep-2009/msg00269.html

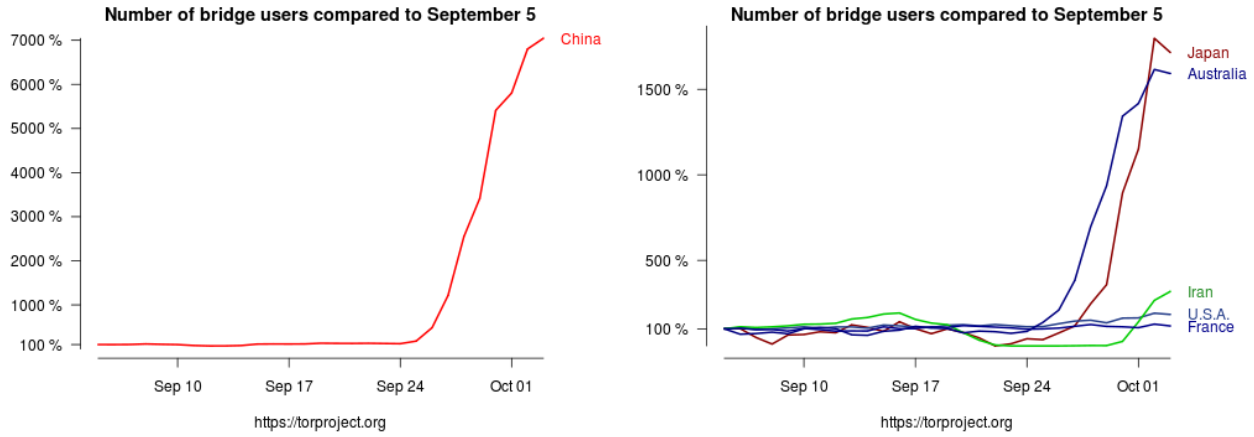

Figure 1: Possibly false classification of Chinese IP addresses as Australian or Japanese addresses

2 Data basis

We have included the following four databases in the evaluation:

1. the June 2009 database from http://ip-to-country.webhosting.info/ as it is included in current Tor versions,
2. the most recent (as of October 19, 2009) freely available Maxmind database from http://www.maxmind.com/,
3. the most recent (as of October 19, 2009) GNUv3-licensed database from http://software77.net/geo-ip/, and
4. a copy of the commercial Maxmind database from October 20, 2009.

3 Coverage of databases

We start with comparing what IP address ranges the GeoIP databases cover. We iterate over all 2^{32} possible IPv4 addresses and visualize what country codes these addresses evaluate to. The comparison result is an image of $1\,024 \times 1\,024$ pixels with the color of each pixel showing what country code the addresses resolve to. Every pixel in this image represents $4\,096 = 2^{12}$ IP addresses, so that all $1\,024 \times 1\,024 = 2^{10} \times 2^{10}$ pixels display all $4\,294\,967\,296 = 2^{32}$ possible IP addresses. The first line of the image contains IP addresses 0.0.0.0 to 0.63.255.255, the second line 0.64.0.0 to 0.127.255.255, etc. The colors are chosen rather arbitrarily just in order to distinguish different ranges: country codes from AA to IQ are displayed in different shades of red, codes from IR to RH in shades of green, and codes from RI to ZZ in shades of blue.

In this and all subsequent analyses, the following ranges of reserved IP addresses have been removed from all databases: 0.0.0.0/8, 10.0.0.0/8, 127.0.0.0/8, 169.254.0.0/16, 172.16.0.0/20, 192.0.0.0/24, 192.0.2.0/24, 192.88.99.0/24, 192.168.0.0/16, 198.18.0.0/16, 198.51.100.0/24, 203.0.113.0/24, and 224.0.0.0/3. We further removed non-country codes like A1 (Anonymous Proxy) and A2 (Satellite Provider) from both Maxmind databases before the analysis.

Figure 2 shows the visualizations of the four databases. These images show the general distribution of assigned IP addresses with the reserved ranges being blank. However, the differences between the databases seem to be negligible on these images.

4 Pairwise comparison of databases

In the next step towards evaluating accuracy of the various GeoIP databases, we perform a pair-wise comparison. Obviously, this comparison cannot show which database is more accurate than the other, but we might be able to detect artifacts by using this approach.

The evaluation iterates over all 2^{32} possible IPv4 addresses and compares the results of the two databases. There are four possible cases for this comparison:

1. only the first database resolves the IP address to a country code,
2. only the second database resolves the IP address to a country code,
3. both databases resolve the IP address to two different country codes, or
4. both databases resolve the IP address to the same country code.

We visualize the difference between the two databases for the number of cases 1 to 3. The comparison result is, again, an image of $1\,024 \times 1\,024$ pixels with the color of each pixel showing how different the two databases are. The pixel color visualizes the fraction of cases (1 to 3) that could be observed when comparing the two compared databases: addresses that are only contained in the first database (case 1) are displayed in different shades of red; IP addresses that were only found in the second database (case 2) are displayed in shades of blue; addresses having different country results in both databases (case 3) are displayed in shades of green; addresses that are resolved to the same country (case 4) do not add any color to the pixel. For example, a full line (or even area) of red means that only the first database contains a resolution for the IP addresses in the given range.

Figures 3 shows the comparison of the Tor database with the three other databases as well as the comparison of free and commercial Maxmind databases. The comparison of the Tor database with both the free and the commercial Maxmind database shows a large number of red and blue lines indicating addresses are only contained in either of the databases. In addition to that, there are some green lines showing that the databases disagree on the country resolution.

The comparison of the Tor database with the Software 77 database shows a rather different picture. These two databases disagree in many more places, as shown by the large green areas. The image also shows some artifacts in the middle of the image. Many of the green lines are exactly 65 536 IP addresses long, which corresponds to a /16 network. These lines are not visible in the comparison to the Maxmind databases. It might be that the Software 77 database has a much lower resolution than the other databases.

The comparison of the free with the commercial Maxmind database shows only very few differences which are mostly red and blue lines. This means that the two databases cover slightly different IP address ranges, but in general they are very similar.

5 Work left to do

- Which are the official reserved address ranges? The ones listed here have been taken from the Software 77 database.
- Do we need to handle region codes like `AP` (Asia/Pacific Region) and `EU` (Europe) in a special way?
- In the next step, focus only on possibly censoring countries that are interesting to Tor: Azerbaijan (AZ), Belarus (BY), China (CN), Egypt (EG), Iran (IR), Jordan (JO), Kazakhstan (KZ), Morocco (MA), Myanmar (MM), Pakistan (PK), Russia (RU), Saudi Arabia (SA), Sudan (SD), Syria (SY), Tunisia (TN), U.A.E. (AE), Uzbekistan (UZ), Viet Nam (VN), and Yemen (YE). As a possible (though not perfect) metric: how many IP addresses do the GeoIP addresses resolve to these countries? The more, the better?
- Try confirming/falsifying samples of resolutions by making requests to the WHOIS database or using some other networking fu.
- Try to learn what changes in the regular updates: are those only new assignments, or are existing ranges re-assigned to other countries, maybe even following a pattern?

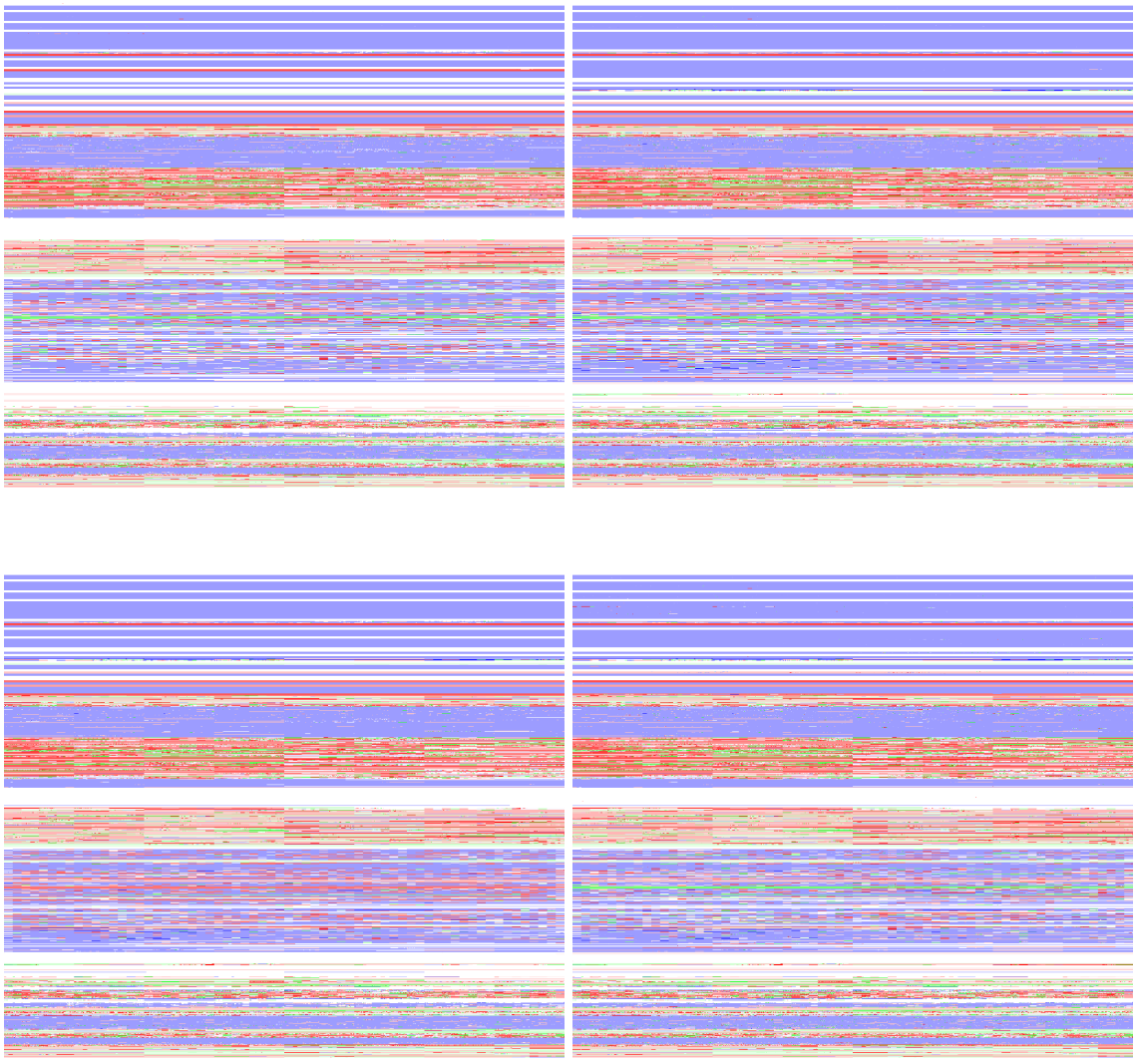

Figure 2: IP address coverages of Tor database (top left), free Maxmind database (top right), Software 77 database (bottom left), and commercial Maxmind database (bottom right)

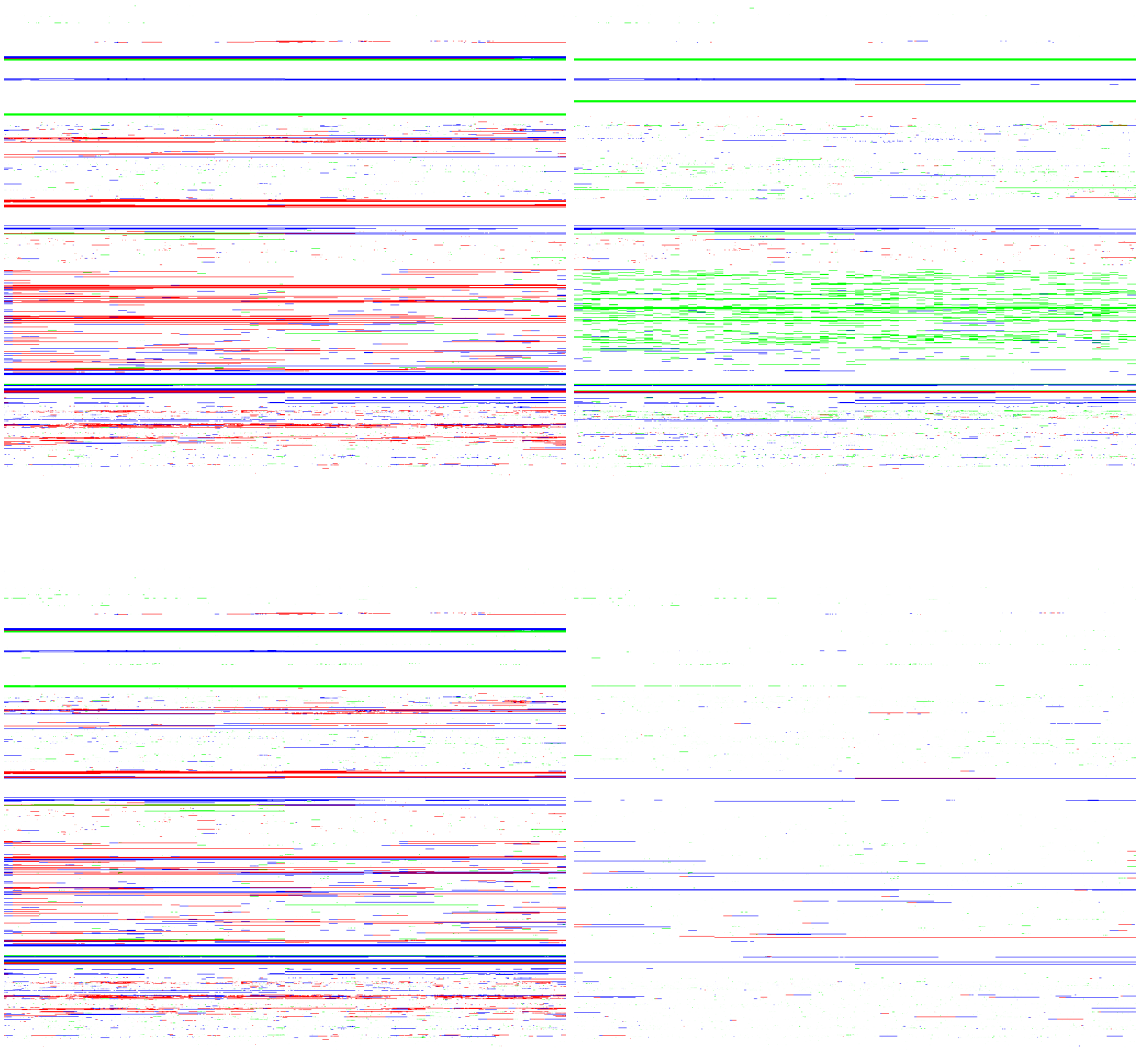

Figure 3: Comparison of the Tor database with the free Maxmind database (top left), with the Software 77 database (top right), with the commercial Maxmind database (bottom left), and comparison of free with commercial Maxmind database (bottom right); red = only in first database, blue = only in second database, green = different results