What fraction of our bridges are not reporting usage statistics? (DRAFT)

Karsten Loesing

April 25, 2012

1 Introduction

Tor's current approach to count daily bridge users is probably broken. The estimate of daily bridge users from all countries ranges between a few hundred to half a million in the time between mid-2008 and early 2012 (see Figure 1). We have little idea whether the real number is closer to the lower or the upper end. It's probably "somewhere in the middle."

The current approach to estimate the number of bridge users is based on bridges reporting the number of unique IP addresses they see in a given 24-hour timeframe to the bridge authority. We collect all reports, sum up unique IP addresses per day, and interpret the result as estimated user number.

We already identified two shortcomings in this approach [1]: The first shortcoming is that the assumption that a bridge user only connects to a single bridge is very likely false. As a result we may over-count bridge clients connecting to two or more bridges. The second shortcoming is that we're excluding a yet unknown fraction of bridges which don't report usage statistics to the bridge authority. A possible reason for not reporting statistics is an uptime of less than 24 hours which is the minimum time for reporting statistics to hide exact connection times and to protect the users' privacy.

In this report we want to focus on the second shortcoming by analyzing what fraction of bridges are not reporting usage statistics. Obviously, whether this fraction is at 20% or at 80% has a major impact on the estimated number of bridge users. But in addition to that we hope to learn something more general about how bridges report statistics to the bridge authority that we can apply to new approaches that estimate daily bridge users.

In the following we discuss reasons for discarding reported bridge statistics and possible causes for bridges not to report statistics. We then look into the bridge descriptor archives to quantify what fraction of bridges are affected by these cases. We conclude with ideas for increasing the fraction of included statistics.

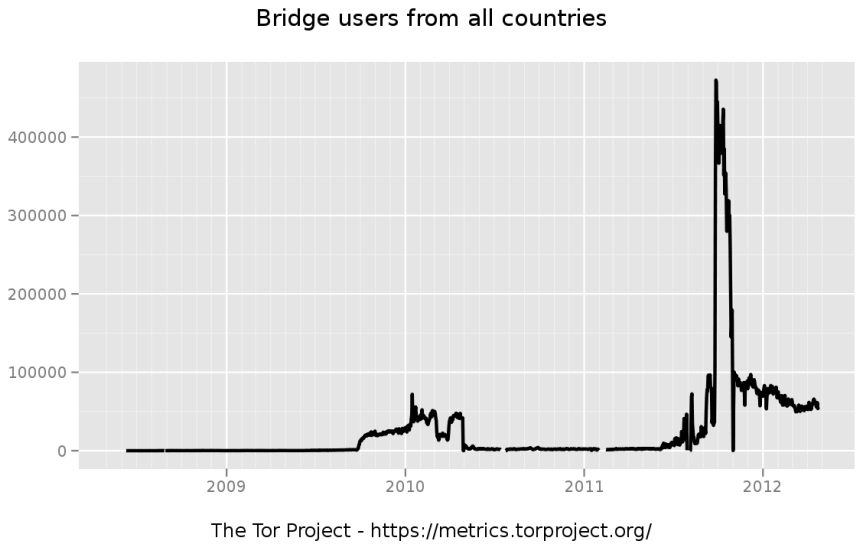

Figure 1: Estimated bridge users from all countries between 2008 and 2012.

2 Reasons for missing bridge usage statistics

There are two categories of reasons for missing bridge usage statistics: either a bridge reports statistics which are discarded, or the bridge does not report statistics at all. Reasons for discarding reported statistics are:

1. **Running as non-bridge relay:** We exclude all statistics from bridges that have been running as non-bridge relays before. The reason is that non-bridge clients may still connect to such a bridge. We expect there to be many more directly connecting users than bridge users, so including these statistics might lead to greatly overestimating the number of bridge users. We currently exclude statistics from bridges which have been running as relay at *any* time in the past, even months ago. We had cases where excluding such a bridge removed a sudden increase in bridge user numbers which could not be explained otherwise.

2. **Known bug in statistics code:** There are a few Tor versions which had bugs in their statistics implementation. We exclude these statistics, too.

3. **Missing geoip file:** We recently discovered that bridges which don't have a geoip file still report bridge usage statistics with all zeros. For the current approach where we sum up all observations, this isn't a problem. But it's still interesting to learn how wide-spread the problem of missing geoip files on bridges is. Only bridges running Tor version 0.2.3.1-alpha or higher report whether they have a geoip file configured or not.

In addition to these cases, there are a few possible causes for bridges not reporting statistics:

4. **Less than 24 hours uptime:** Bridges which have an uptime of less than 24 hours don't report statistics for this period of time. This has to do with the requirement to aggregate observations for a sufficient amount of time to hide exact connection times and to protect the users' privacy.

5. **Descriptor publication delay:** Some bridges may even complete a 24-hour interval and prepare statistics to be reported in their next descriptor. But then they go offline and don't publish that descriptor. Bridges look at previously finished statistics intervals when starting up, but either a bridge decides that its previous statistics are too old to be published, or a bridge never shows up again. The fix here might be to make bridges publish a new descriptor immediately after finishing a statistics interval, which is suggested as enhancement #4142. We should probably find out how many bridges are affected by this problem before implementing the fix.

6. **Other reasons:** There may be other causes for a bridge not reporting statistics which we did not identify.

3 Fraction of missing bridge usage statistics

After listing reasons for reported observations being discarded and for bridges not reporting statistics at all, we now want to quantify how many bridges are affected by which case.

Figure 2 shows the fraction of bridges that did or did not report usage statistics and how many of these reports had to be discarded. The graph shows an almost monotonic downward trend of non-reported statistics from 2008 to early 2010 to around 20%. This fraction went up only slightly in 2010 and 2011 to 40% and is now back at 20%. The fraction of reported and discarded statistics was between 10% and 20% for most of the time between 2008 and today. As a result, the fraction of reported and used statistics went up from around 35% in early 2009 to around 75% in early 2012.

These results are much better than expected before. A bridge usage statistic that is based on 75% of all bridges at least rules out inaccuracies from too little sample sizes. We still want to look into reasons for discarded or not reported statistics.

Figure 3 shows what fractions of reported bridge statistics were discarded for what reasons. The fractions of statistics that had to be discarded because of the geoip-stats bug in Tor 0.2.2.x or because of missing geoip files are at almost 0% for most of the time. Only in late 2009, the geoip-stats bug affected up to 5% of bridges. But the fraction of discarded statistics because of bridges previously running as non-bridge relays is quite high at 10% to 20%.

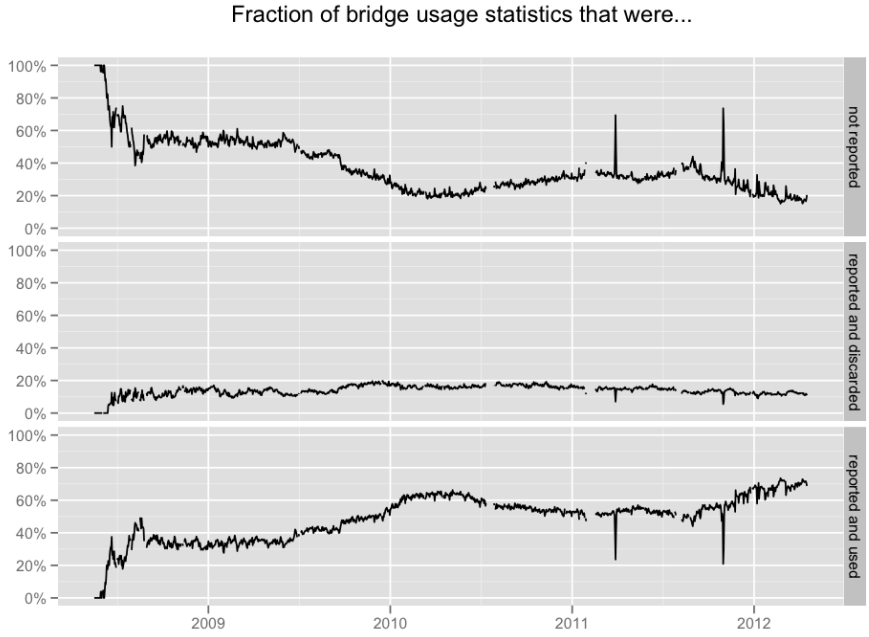

Figure 2: Fraction of bridges that reported statistics which were either used or discarded, or that did not report statistics.

It's quite likely that we could reduce this fraction by being less strict about bridges running as non-bridge relays. In theory, a delay of a few days between running as relay and running as bridge should be sufficient to exclude directly connecting clients from the statistics. However, this requires further analysis.

Figure 4 shows what fractions of bridges don't report statistics for what reasons. For most of the time, the fraction of bridges not reporting statistics because they went offline before their 24-hour interval ended was about as large as the fraction of "other reasons." Only recently, the "other reasons" dropped to almost 0%. The fraction of missing statistics due to a delay between completing a statistics interval and publishing the descriptor containing those statistics is almost at 0% for most of the time.

There's probably not much that we can do about the first category of bridges which go offline before their 24-hour interval ends. This interval is there to hide exact connection times and to protect the users' privacy. Any algorithm will have to cope with 15% to 25% missing statistics due to the 24-hour interval requirement. Fortunately, the fraction of non-reported statistics due to the descriptor publication delay is almost at 0%, so we don't have to fix that. It's unclear what other reasons led to bridges not publishing statistics. Given that this fraction is almost at 0%, there's no immediate need to investigate.

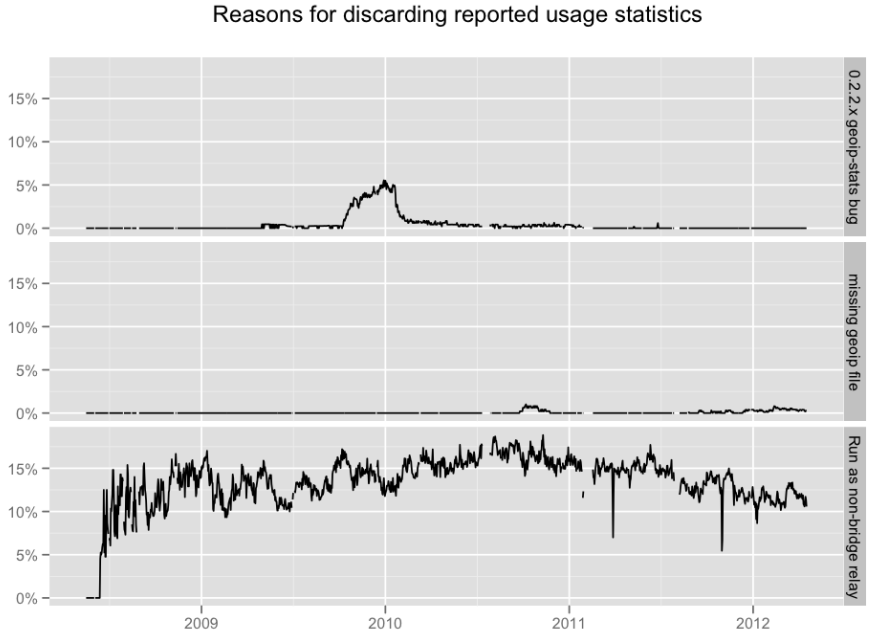

Figure 3: Reasons for discarding reported usage statistics.

4 Conclusion

In this report we analyzed what fraction of bridges are not reporting usage statistics, which might affect our daily bridge user estimates. The analysis of bridge descriptor archives resulted in a fraction of up to 75% of bridges reporting usage statistics that get used to estimate user numbers. This fraction might even be increased by discarding fewer statistics from bridges that were seen as non-bridge relays before. We conclude that a too small sample size is not the issue of our probably wrong bridge user numbers. We think that a new approach that will be based on bridges reporting their findings in 24-hour intervals has a good chance of leading to quite reliable user numbers.

References

[1] Sebastian Hahn and Karsten Loesing. Privacy-preserving ways to estimate the number of Tor users. Technical Report 2010-0001, The Tor Project, November 2010.

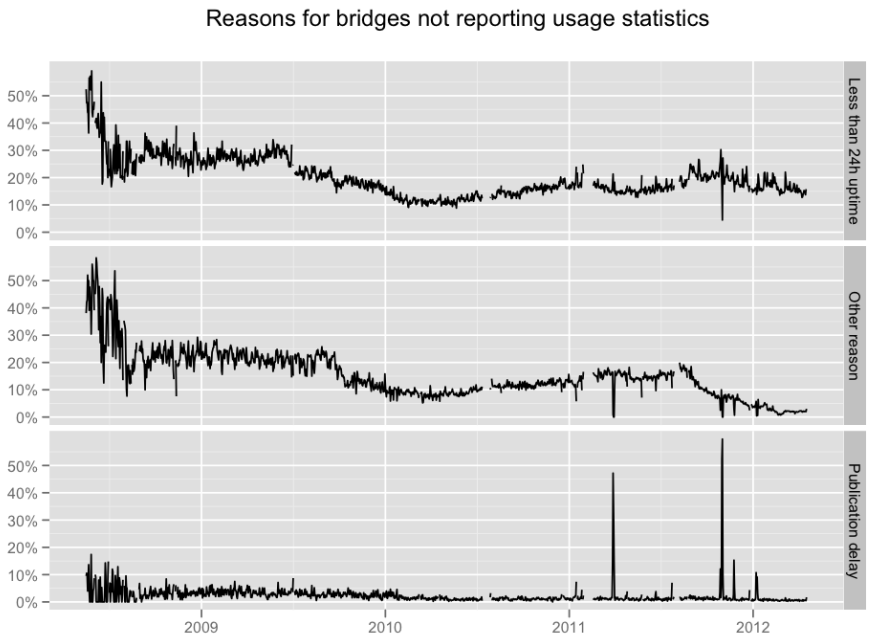

Figure 4: Reasons for bridges not reporting usage statistics.