Measuring the Tor Network

Evaluation of Client Requests to the Directories

Karsten Loesing

June 25, 2009

Abstract

Only few facts are known about usage of the Tor network. The number of daily users in the Tor network is still subject to educated guesses, and there are only few data available on the distribution of users to countries. This report analyzes client requests to twelve directories measured over 4 weeks in June 2009. Results include an estimation of the total number of Tor users and their distribution to countries.

1 Motivation

While a few facts are known about the infrastructure of the Tor network, we are still lacking many facts about its usage. The two most important questions that come to mind are:

- How many users does the network have?
- Where do these users come from?

This report makes an attempt to answer these two questions by analyzing client requests to the directories. The rationale is that directories obtain a local view on the network from the number of connecting clients and the number of requests they receive. We propose a formula to derive a possible global view on the network.

Further, the directories break down their observations of clients and requests by country by using a GeoIP database. From these data we can tell what fraction of users comes from which countries.

2 Data basis

This report is based on the measured client requests to a set of twelve directories (four directory authorities and eight directory mirrors). These twelve directories

Table 1: Directories measuring GeoIP statistics for this analysis

	Configured Bandwidth (KiB/s)			
Nickname	Rate	Burst	MaxAdvertised	Operator
trusted	13312	15360	–	Jacob Appelbaum
badbits	20480	51200	–	Jacob Appelbaum
moria1	–	–	10	Roger Dingledine
moria2	–	–	20	Roger Dingledine
moria5	50	1000	–	Roger Dingledine
xpdmTindome	50	200	20	Marcus Griep
fluxe3	150	200	–	Sebastian Hahn
gabelmoo	1024	1500	500	Karsten Loesing
hamsterrad	200	500	–	Karsten Loesing
ephemer2	90	–	–	Steven J. Murdoch
ides	–	–	12	Mike Perry
vallenator	2100	4000	–	Hans Schnehl

have been instrumented[1] to measure client requests and write them to a local file every 24 hours. These measurements have been performed between May 28 and June 25, 2009 with some directories being started later, being restarted (and therefore losing observations of the current 24-hour interval), or ending their measurements earlier. Table 1 contains a list of these twelve directories together with their bandwidth settings.

The data that each of the twelve directories writes down after 24 hours of measurements consists of three main parts (see Figure 1 for an example):

- Unique IP addresses: The directories memorize which IP addresses they have seen within the past 24 hours and output the number of unique addresses per country (`ns-ips` and `ns-v2-ips` lines).
- Directory requests: The directories also count the total number of requests by country, regardless of whether they come from an already known or a new IP address (`n-ns-reqs` and `n-v2-ns-reqs` lines).
- Share of requests: The directories determine what share of requests they should see based on the probability of clients to pick them rather than other directories (`v2-ns-share` and `v3-ns-share` lines).

Unique IP addresses The directories count the number of unique IP addresses by country that they have seen over the past 24 hours. For the sake

[1]More information on setting up a directory to measure these data can be found here: `http://archives.seul.org/or/dev/Jun-2009/msg00000.html`

```
written 2009-05-28 18:53:15
started-at 2009-05-27 18:53:00
ns-ips us=1056,de=536,fr=360,cn=208,kr=176,it=160,gb=152,..
ns-v2-ips us=808,de=408,cn=296,kr=144,gb=136,ca=128,fr=104,..
requests-start 2009-05-27 18:53:00
n-ns-reqs us=1152,de=552,fr=376,cn=232,kr=232,gb=160,it=160,..
n-v2-ns-reqs us=888,de=424,cn=320,kr=240,gb=144,ca=136,fr=112,..
v2-ns-share 0.25%
v3-ns-share 0.26%
```

Figure 1: Example data of directory requests measured over 24 hours

of simplicity, every IP address is assumed to belong to a single user in the following analysis.[2] Figure 2 shows the number of unique IP addresses that the directories have seen in 24-hour intervals. The large differences in number of IP addresses are the result of different probabilities for clients selecting the directories. Directory mirrors see more requests the more bandwidth they advertise, which is the minimum of the bandwidth rate (`Rate` column in Table 1), the maximum advertised bandwidth (`MaxAdvertised` column), and the maximum observed bandwidth (not shown in the table, varies over time). In this graph, IP addresses requesting both versions of network statuses are counted twice, as the only available data are the numbers of unique IP addresses requesting a certain network status version. This simplification seems acceptable, as clients do not request both network status versions during normal operation.

The numbers are stable for most of the directories, except for `vallenator` that exhibits decreasing numbers over time. The reason is that the longer this directory ran, the more clients connected to it and the more version 2 network status requests were rejected with a `503 Busy` response. These unsuccessful requests (and the corresponding IP addresses) are not counted in the statistics.

Further, there is a sudden decrease in the number of IP addresses seen at `hamsterrad` on June 9. This is the time when the relay obtains the `Guard` flag for the first time. Clients weight the probability of picking a guard node as directory mirror with only one third of the original probability.

Similarly, `badbits` sees a rather low number of IP addresses compared to its advertised bandwidth due to the fact that it has the `Exit` flag.

Directory requests The directories further count the number of requests per country. Figure 3 shows total request numbers as the sum of requests for version 2 statuses and version 3 consensuses. Obviously, these numbers are higher than the number of unique IP addresses, because the same IP address can request more than one network status from the same directory within 24 hours.

The number of requests seen by the four directory authorities `gabelmoo`, `ides`, `moria1`, and `moria2` are much higher than they would be when these four

[2]This assumption may be wrong with users being connected via dynamic IP addresses or using network address translation.

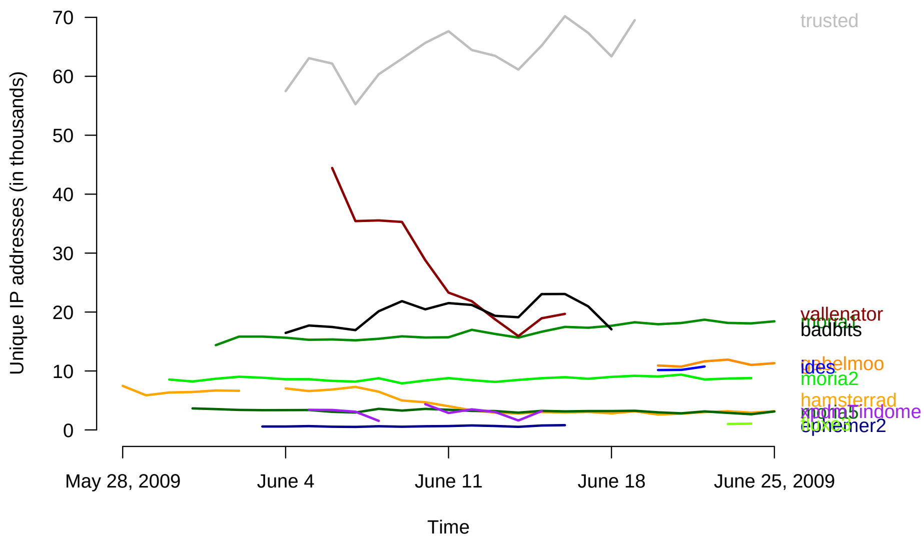

Figure 2: Number of unique IP addresses

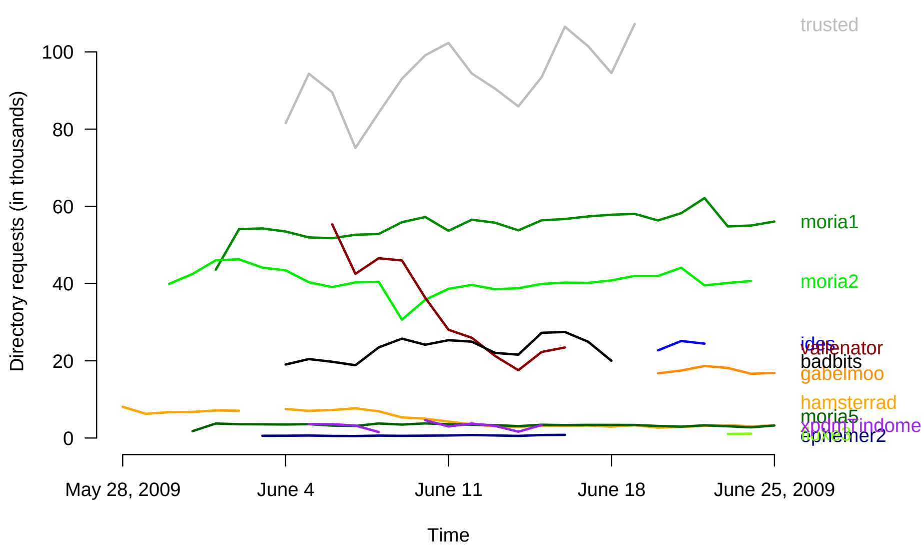

Figure 3: Total number of requests

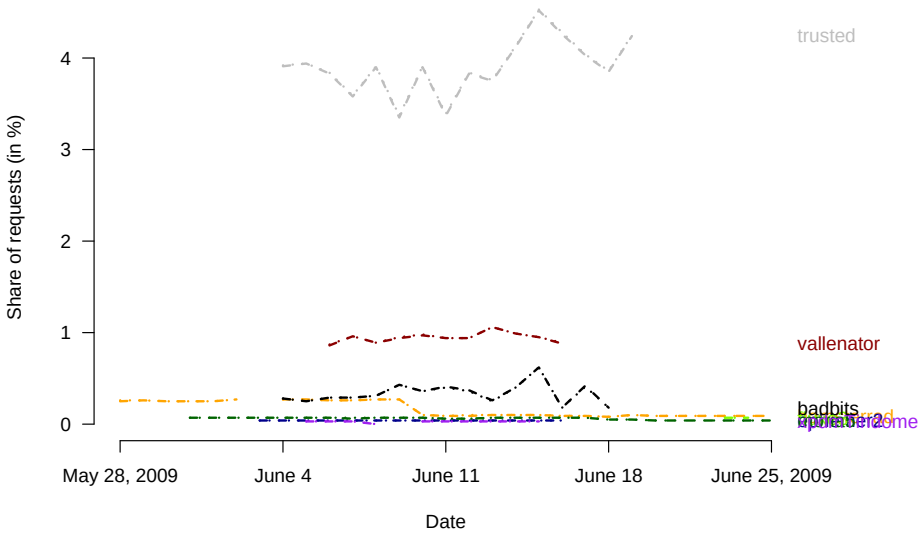

Figure 4: Shares of directory requests that directories think they should see

were directory mirrors. The reason is that bootstrapping clients only know the addresses of the authorities and need to fetch their first network status from them.

Share of requests The third kind of data that directories report every 24 hours is the share of requests they think they should see. Figure 4 shows these shares for both version 2 network statuses (dashed lines) and version 3 consensuses (dotted lines). In most cases (without visible exception in the graph) these two shares are identical.

These shares are based on the directories' own advertised bandwidth as compared to the total advertised bandwidth in the network. These numbers do not take into account that directories might fail or are busy and therefore deny client requests. As a result, the real share that a directory sees would be higher than expected, because clients that fail at one directory retry at another subsequently.

Table 2 shows the (non-representative) results of one test run to download v3 consensuses from all directory mirrors in the network performed on June 24. Requests were started with a delay of 10 seconds and given 10 minutes to finish. Of the 875 requests, only 309 (35%) succeeded with status code 200. The other 65% were considered as bad request (status code 400, 0.1%), were not found (status code 404, 4%), were rejected because the directory was busy (status code 503, 36%), or failed because no connection could be established to the directory (labelled Error in the table, 24%). Connection errors included unability to find a route to the host, connection refusals, and timeouts.

Table 2: Results of version 3 consensus downloads from all directory mirrors

	200	400	404	503	Error	All
Number of requests	309	1	35	318	212	875
Bandwidth sum (KiB/s)	233077	253	7666	39302	13650	293948
Bandwidth mean (KiB/s)	754	253	219	124	64	336

When considering the advertised bandwidths of directories, the total bandwidth of these directories answering with status code 200 is 233077 KiB/s (79%) and therefore much higher than only 35%. That means that 79% of all client requests are answered correctly. The mean bandwidth of directories accepting directory requests is 754 KiB/s in contrast to 253, 219, 124, or 64 KiB/s for failing directories for the various reasons. Apparently, the likelihood of a positive answer increases with the advertised bandwidth of a directory. As a result, all reported shares in this report are divided by 80% to compensate failing directory requests in the network.

3 Estimating total user numbers

The most important metric to be answered by this report is the number of users that connect to the network per day. This report makes an attempt to estimate total user numbers, focusing on version 3 network consensuses, i.e., on client versions 0.2.0.x or higher.[3]

The directories measuring directory requests each have only a local view on the network. In the following analysis, these local views shall be used to derive a global view of the number of users in the network. In the following, three attempts are made to estimate the number of users: First, the number of new users are estimated from the number of requests seen at the directory authorities. Second, we guess the number of regular users from the requests seen at directory mirrors under the conservative assumption that every user makes 10 requests for network statuses per day. And third, we try to derive a way to estimate for the number of regular users more accurately which, however, still appears to be faulty.

Estimate of new users The number of new users can be estimated from looking at the number of requests that the directory authorities see. New clients do not know any directory mirrors and therefore have to ask one of the directory authorities for the current network status. Clients are selecting one of

[3]It is significantly harder to estimate the number of clients running versions 0.1.2.x or older. The reason is that many requests for version 2 network statuses are rejected with a `503 Busy` reply, especially on directories with rather low bandwidth. This adds considerable uncertainty into estimates. However, with 0.1.2.x being phased out, the fraction of clients downloading version 2 will soon decrease anyway.

Table 3: Shares of directory requests coming from new users that directory authorities should see

	`gabelmoo`	`ides` and `moria1`
May 28 to June 11	$\frac{56\%}{56\%+5\times100\%} = 10.072\%$	$\frac{100\%}{56\%+5\times100\%} = 17.986\%$
June 12 to June 25	$\frac{56\%}{56\%+4\times100\%} = 12.281\%$	$\frac{100\%}{56\%+4\times100\%} = 21.930\%$

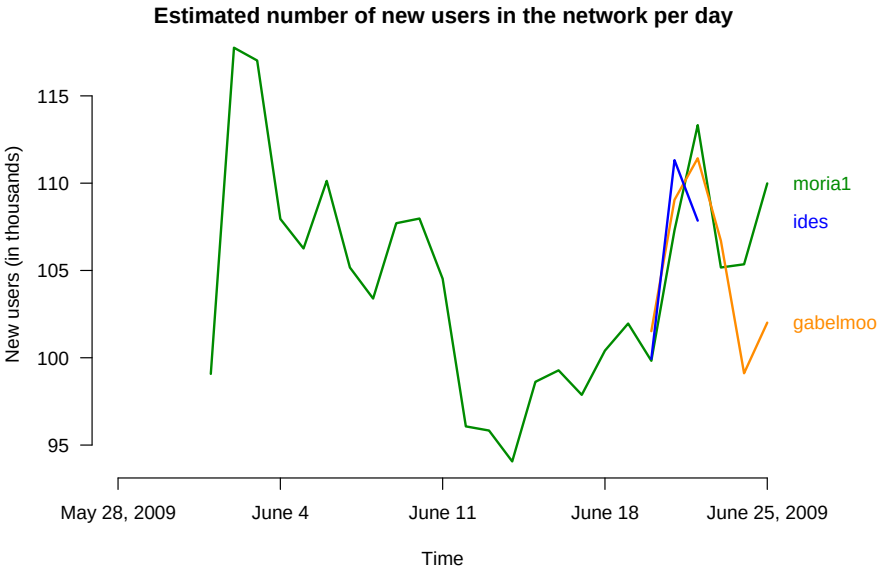

Figure 5: Estimate of new users in the network

the currently six authorities at random with equal probability. However, that does not necessarily mean that every authority sees exactly 1/6 of all requests: The authority `dannenberg` has been offline since June 11, 2009, 14:00 UTC, so that the other authorities have processed 1/5 of all requests. Further, the IP address of `gabelmoo` has changed in December 2008, so that some clients tried to download the consensus from the old IP address and failed. Between June 20 and 22, `gabelmoo` has received about 56% as many requests as `moria1` and `ides` did. For this analysis we use the factors from in Table 3 as estimates to conclude the number of new users from local observations. Figure 5 shows the resulting estimate of new users.

Conservative estimate of regular users As soon as clients have bootstrapped, they need fresh network statuses every 2 to 3 hours. Clients download these network statuses from directory mirrors and avoid bothering the directory authorities again. As a result, we should be able to count the number of regular

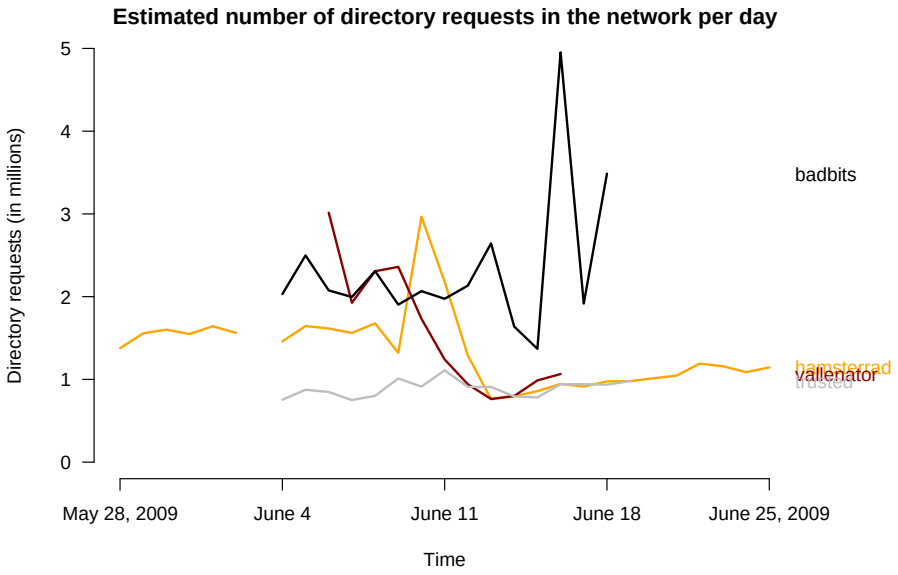

Figure 6: Estimate of total request numbers in the network

users from the local views that the directory mirrors have. In the following analysis, only those directory mirrors seeing at least 0.1% of all requests shall be considered. The three data points for this estimation are:

- Number of locally seen requests: r
- Number of directly connecting clients: c
- The corrected share of requests that we should see: s

From these numbers we can determine the number of *requests* in the whole network (Q): $Q = r/s$. The assumption is that every client makes an independent decision for each network status download which directory to ask. Figure 6 shows estimates of total requests in the network per day.

The numbers of requests in the network (Q) is the product of the number of clients (N) and the number of requests each client sends on average (x). There are probably very different usage patterns influencing the average number of requests that each user sends per day. Some users might connect only for a few minutes while others are connected to the network for the whole day. The former users would send exactly 1 request per day, the latter would send a new request every 2–3 hours, i.e., up to 10 requests per day. For this conservative estimation of user numbers we make the assumption that every user requests 10 network statuses per day. This assumption probably leads to undercounting the number of users. Figure 7 shows the estimated number of clients using this assumption. Most estimates are in an interval from 100,000 and 300,000 users per day, however closer to 100,000 than to 300,000.

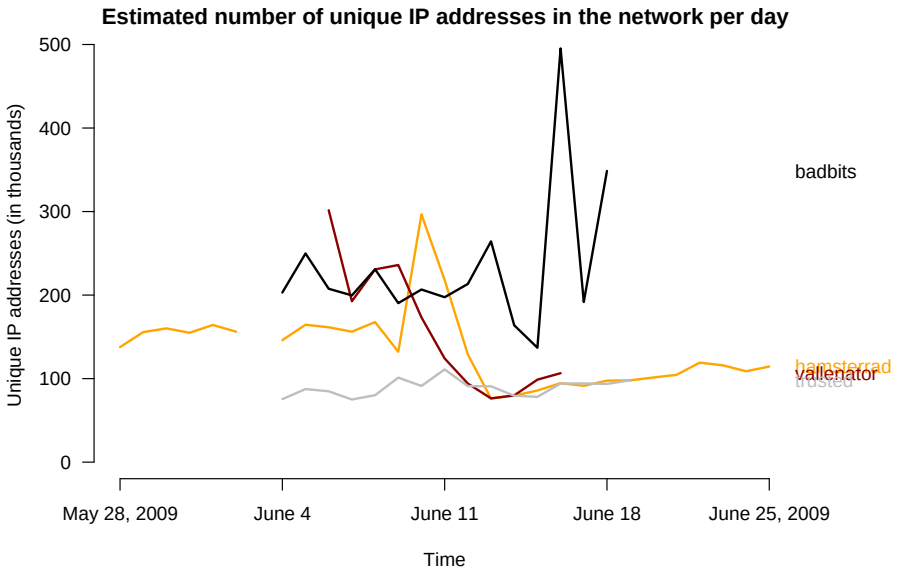

Figure 7: Conservative estimate of total users in the network

Experimental estimate of regular users Instead of making assumptions for the number of requests that each user makes, we can try to derive this number from local information. Therefore, we take advantage of two relationships between observed numbers:

- We already know that the total number of requests (Q) is the product of every user (N) sending a certain number of requests on average per day (x): $Q = N \times x$.
- From the assumed average number of requests (x) and the share of requests that we should see (s), we can determine the probability for a client to ask us *at least once* in 24 hours. This probability is the complementary probability of not being asked a single time: $1 - (1 - s)^x$. We can use this probability to estimate how many clients there are in the network from the number of clients we have seen: $N \times (1 - (1 - s)^x) = c$.

These two equations with 2 variables (N and x) can be combined to one equation with 1 variable (N): $N \times (1 - (1 - s)^{Q/N}) = c$. In the next step, we can calculate N and x separately.

Figure 8 shows the estimates of requests per clients and Figure 9 the estimates of total users in the network.

Unfortunately, these results are very likely wrong, as an average of 40 requests per client per day seems highly unrealistic. Finding the error in this calculation is subject to future analyses.

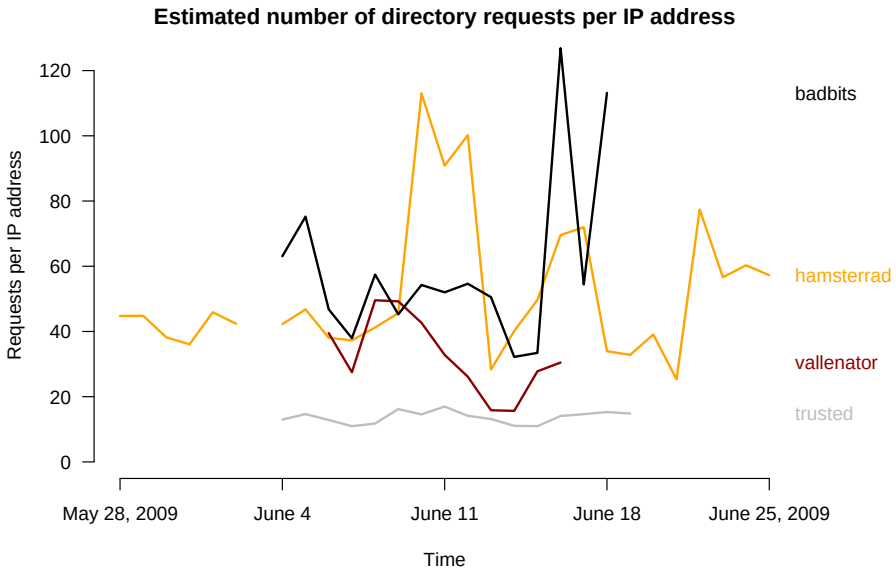

Figure 8: Experimental estimate of requests sent per users

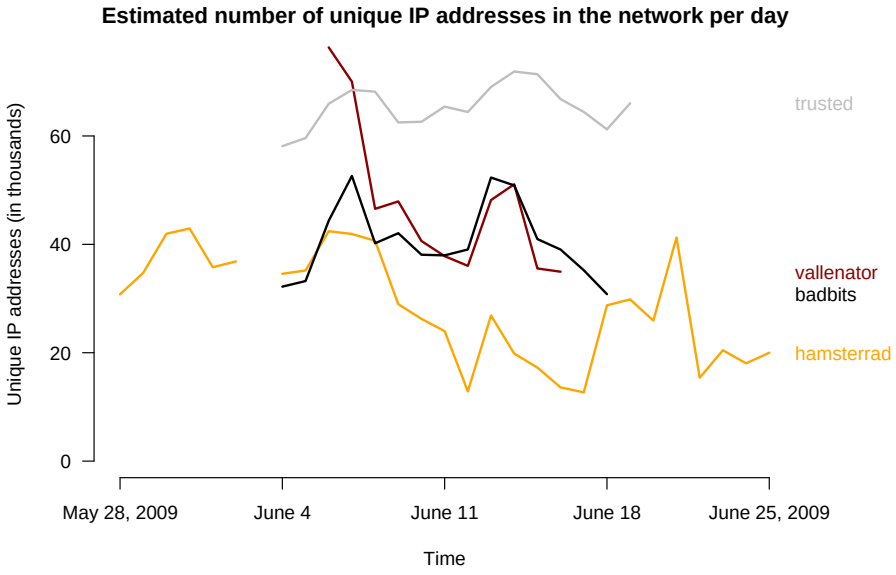

Figure 9: Experimental estimate of total users in the network

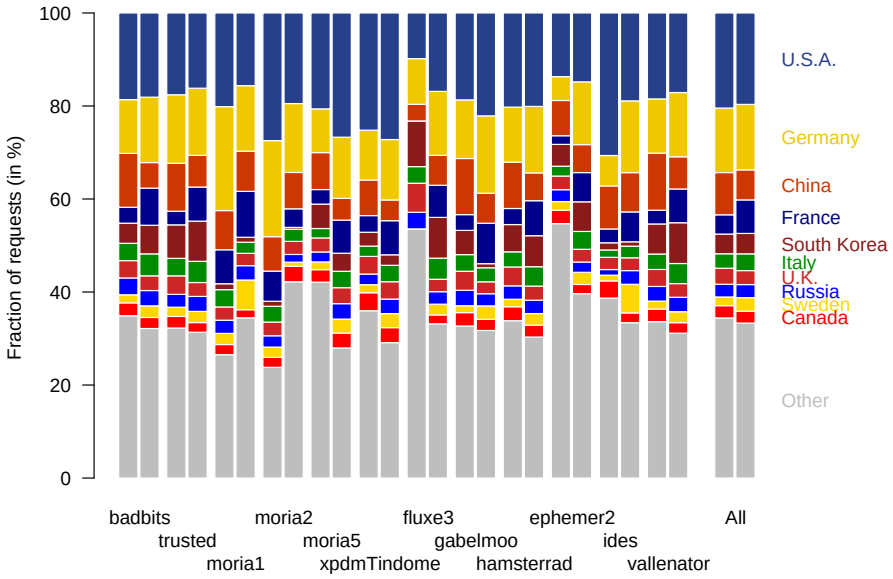

Figure 10: Fractions of users by country

4 Number of users by country

The next interesting metric is the distribution of users to countries. These numbers can be determined more easily, as the directories already break down their observations by country.

Figure 10 shows the fractions of users by country as seen on the twelve directories and averaged over all of them. The left bars denote version 2 requests, the right bars version 3 requests. Only the top 10 countries are displayed.

Figure 11 shows the same fractions for 19 countries which might restrict Internet usage for local users. In addition to the named seven countries, the graph also shows client requests (in decreasing order) from Kazakhstan, Belarus, Jordan, Syria, Yemen, Azerbaijan, Uzbekistan, Myanmar, Egypt, Morocco, Sudan, and Tunisia. One interesting point in this figure is the large share of Iranian version 3 requests answered by `gabelmoo` in comparison to the other two authorities `moria1` and `ides`. It is noteworthy in this context that `gabelmoo` is the only authority of these three listening on directory port `443`.

5 Future work

One of the next steps in this analysis is to find the mistake in the experimental estimation of user numbers. The assumption of 10 requests per user per day seems too high, though we are still missing a better number. The result of the experimental estimation of 40 requests per user per day seems even less realistic, though.

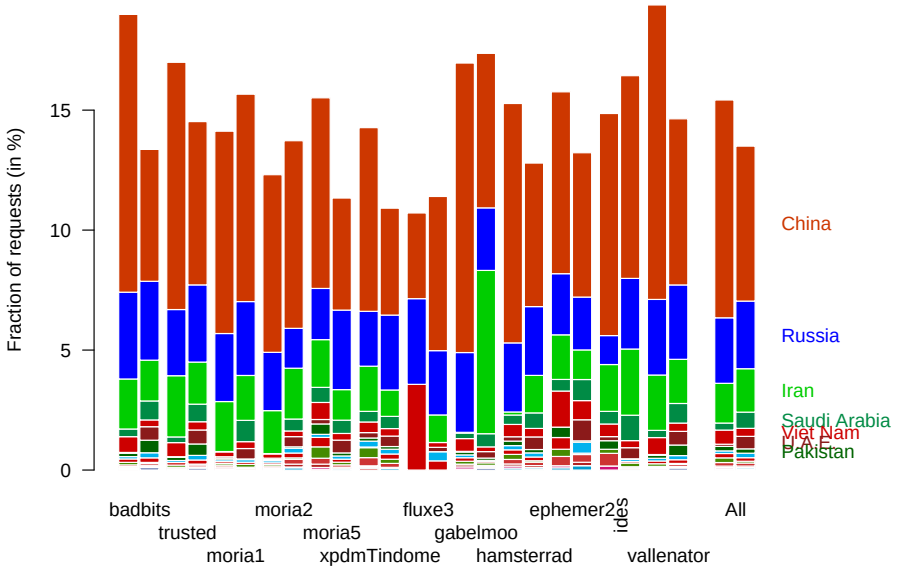

Figure 11: Fractions of users in potentially censoring countries

This analysis is based on the most comprehensive data set available on network usage up to this point. The logical next step is to measure and aggregate directory requests on most or all directories in the network and establish a central repository for these aggregate data. It seems that directories with configured bandwidths of at least 200 KiB/s would be most useful.

Another direction for future work is the comparison of requests to the directories with bridge clients connecting to bridges or regular clients connecting to entry guards. Bridges already gather similar statistics about their users, and it is planned to make entry guards do the same in the near future. While entry guards should see similar total user numbers and distributions to countries, bridges might exhibit a different user set.