Measuring the Tor Network from Public Directory Information

Karsten Loesing
karsten@torproject.org

Tor Tech Report 2009-08-002*
August 7, 2009

Abstract

The Tor network is one of the largest deployed anonymity networks, with more than one thousand servers, called relays, and probably hundreds of thousands of clients. A few facts are known about the Tor network, even though single individuals or organizations, including The Tor Project, control only small parts of the network. The reason is that clients need to know which relays there are in the network in order to build circuits and anonymize their traffic. Therefore, relays periodically report their addresses and capabilities to a small set of directory servers which redistribute relay lists as directories to clients. Everyone can collect these directories and analyze them to obtain statistics about the relays in the Tor network. In this paper we demonstrate statistics about the Tor network from February 2006 to February 2009 solely based on archived directory information. The goal is to observe trends in the network without having to collect any data that might compromise the security or anonymity of Tor users.

1 Introduction

Privacy-enhancing technologies have become an invaluable tool to help people protect their privacy on the Internet. Privacy-enhancing technologies allow people to stay anonymous or pseudonymous by redirecting their traffic over multiple nodes in an anonymity network. Tor [3] is one example for such an anonymity network. The Tor network consists of more than one thousand servers, called relays, which are run on a voluntary basis by individuals and organizations worldwide. These relays are used by presumably a few hundreds of thousands of clients to anonymize their Internet communication. This makes the Tor network one of the largest deployed anonymity networks.

Tor clients obtain anonymity by building circuits of usually three relays that redirect their communication before exiting to the real Internet. All messages along the circuits are encrypted

*This report was presented at 2nd Hot Topics in Privacy Enhancing Technologies (HotPETs 2009), Seattle, WA, USA, August 2009.

multiple times so that neither participant can tell where the circuit begins and where it ends. The result of this redirection is that nobody can link the clients' IP addresses to the data contents. Clients learn about running relays by downloading lists of running relays from directory authorities, a small number of relays that collect and distribute routing information in the network. Clients use this information to decide which relays to select when building circuits.

The Tor network is open in the sense that everyone with a moderately fast Internet connection can join the Tor network and run a relay. The idea is to distribute trust and make it very hard for a single person or organization to control large parts of the network. From the perspective of the Tor project it is desirable to learn about certain characteristics to improve the network for its users. In the present paper we describe an approach to observe trends in the network from evaluating archived directory information. Examples include but are not limited to:

- The directory authorities assign flags to some of the relays that are used by clients to make path selection decisions. The analysis of directory archives can show whether the assumptions for assigning flags are still adequate or need to be changed.

- Knowing about relay versions and platforms in the network can be useful to learn about the update behavior of relay operators and to decide when to request popular operating system distributors to upgrade to a new Tor version series.

- Learning about the number of relays on dynamic IP addresses can help prioritize development efforts to better support IP address changes on relays.

- An analysis of advertised and used bandwidth on relays can give first insights into investigating why Tor is slow and how to improve it.

- Changes in the distribution of relays to countries can give hints that more efforts need to be taken to support relay operators in certain countries.

- Information about the characteristics of the Tor network can help when simulating the effectiveness of attacks or design changes. As an example, the rates of joining and leaving relays per hour were used to derive the average availability of hidden service descriptors in a distributed directory before deployment [5].

This paper focuses only on directory information that is already publicly known. Most of the data is required for the operation of Tor and is downloaded by all clients. All data can be collected without running a single piece of the network infrastructure simply by downloading data from the directory authorities. This approach allows us to gain insights about trends in the network infrastructure and, to a small extent, its usage.

2 Tor Directory Protocol

Relays and clients communicate over the directory protocol [9] to exchange directory information. The reason is that clients need to know which relays exist to build circuits. Clients

Table 1: Subset of the data contained in a router descriptor

Data field	**Description**
Nickname	Name for the relay as chosen by its operator
Address	IP address
OR port	Port that accepts connections from clients and other relays
Dir port	Port that answers directory requests
Average bandwidth	Number of bytes the relay is willing to sustain over long periods
Burst bandwidth	Number of bytes the relay is willing to sustain in short intervals
Observed bandwidth	Estimate of the capacity that the relay can handle
Platform	Tor software version and operating system
Published	Time when the descriptor was generated
Fingerprint	Hash of the relay's public key
Uptime	Number of seconds that the relay has been running
Onion key	Medium-term key used for onion routing
Signing key	Long-term identity key
Exit policy	Rules that define what targets and ports are allowed for exiting
Contact	Address to contact the relay operator
Family	List of relays operated by the same operator

further need to know about the relays' capabilities to make good path selection decisions before building circuits.

The first step in the directory protocol is that relays publish router descriptors to the directory authorities, reporting their current contact data and capabilities. The contact data includes information like the IP address and port to listen for onion routing protocol requests as well as cryptographic keys. Capabilities include, among other things, available bandwidth and exit policy. Table 1 shows the data from router descriptors that can be relevant for statistical analysis.

All data in a router descriptor is signed by the relay and hence cannot be altered by anyone. However, nothing prevents a relay operator from changing the source code, e.g., to lie about the relay's capabilities. Various authors have exploited the fact that relays self-advertise information that clients use for making path selection decisions later to conduct attacks on Tor [1,8].

In addition to router descriptors, relays publish a second type of document to the directory authorities, called extra-info document. Extra-info documents contain data which are not required for normal operation but which might be useful for statistical purposes. Extra-info documents are not meant to be downloaded by clients, but they are publicly available for everyone to download, too. The only data that is contained in extra-info documents right now is the relay's bandwidth history over the past 24 hours. Table 2 shows the two relevant data fields of extra-info documents.

The reason for separating out extra-info documents from router descriptors was that clients should not need to download data that is not essential for their operation. Recent measurements have shown that even with this separation, clients on low-bandwidth access networks spend most of their bootstrapping time on downloading router descriptors from the directories [4]. It

Table 2: Subset of the data contained in an extra-info document

Data field	Description
Read history	Number of bytes received by this relay in 15-minutes intervals during the past 24 hours
Write history	Number of bytes sent by this relay in 15-minutes intervals during the past 24 hours

Table 3: Subset of the data contained in a network status consensus

Data field	Description
Relay identity	Unique relay identity, derived from identity key
Descriptor identifier	Referenced descriptor
Exit flag	Authorities think this relay should be preserved for building exit circuits
Fast flag	Authorities think this relay is suitable for high-bandwidth circuits
Guard flag	Authorities think this relay is suitable for use as entry guard
Stable flag	Authorities think this relay is suitable for long-lived circuits

is desirable for clients with limited connectivitiy to further reduce the size of router descriptors while retaining the data, e.g., for statistical purposes. The recently proposed microdescriptor approach [2] might be an important step into this direction.

The directory authorities store router descriptors and continuously verify availability of relays to maintain a list of running relays. The authorities further assign various flags to each relay based on their knowledge of the whole network to indicate special properties of a relay, e.g., if it is more stable than others. These flags are used by clients to make their path selection decisions. Every hour, the directory authorities exchange their views on the network and agree on a common list of available relays, called network status consensus. Every running relay has an entry in a network status consensus with the data as shown in Table 3.

Clients learn about the available relays by downloading a network status consensus and all referenced router descriptors from the directory authorities.

All the directory information as described here can be easily collected on a regular basis. One could run a relay and configure it to act as a directory cache to obtain this information automatically. Alternatively, one can downloaded all documents from the directories using simple HTTP GET requests. The latter can be automated using scripts.[1]

The directory data are already used to provide an overview of the Tor network. TorStatus is a web-based application to present basic statistics on the Tor network.[2] Only recently, TorStatus was extended by Martin Mulazzani [7] by writing a subset of directory information to a database

[1] See Peter Palfrader's directory archive script that performs this task: https://gitweb.torproject.org/tor.git/tree/HEAD:/contrib/directory-archive

[2] See the project homepage of TorStatus, developed by Joseph B. Kowalski and Kasimir Gabert, at: http://project.torstatus.kgprog.com/trac/ (update August 27, 2012: website does not exist anymore)

and visualizing the collected data in the web interface. Mulazzani's approach differs from the approach taken here by creating an interface for users to let them analyze data about the Tor network rather than collecting data and performing the analysis offline. An integration of both approaches in the future might be beneficial.

3 Results

The following statistics are the result of a continuous collection of directory data from February 2006 to February 2009 by Peter Palfrader. A copy of these data can be requested for research purposes by contacting the author of this paper.[3] The tools to import the directory archives into a database and perform evaluations on them have been made freely available.[4]

Relay flags assigned by directory authorities. The first analysis focuses on the total number of relays and their flags as assigned by the directory authorities. Figure 1 shows the average number of relays per day. The topmost line represents all running relays. The shown flags shall help clients make better path selection decisions rather than picking relays uniformly. Directory authorities assign the `Fast` flag to relays that have at least the advertised bandwidth as 90% of all relays, so that clients can pick these relays for high-bandwidth circuits. The `Stable` flag is assigned to relays that the authorities think are more stable than others and therefore suitable for long-lived circuits. The `Exit` flag indicates that a relay permits at least some connections to exit to the Internet and should therefore be preserved for use as exit relay in a circuit rather than be overloaded by being selected for other positions. The `Guard` flag suggests to clients that a relay is suitable for being selected into a small set of entry guards; the idea to use a fixed set of entry guards is to prevent an adversary from forcing a victim to build new circuits until they control the first relay in the circuit and be able to locate the victim using traffic analysis [8].

The overall trend that can be seen in the graph is that the network grows rather quickly until the beginning of 2008 but starts shrinking from then on. The reason for the shrinking is not immediately visible. The graph also contains a few artifacts that can be explained from external events or from the measurement setup. In the interval from February 2006 to November 2007, the directory authorities did not vote on a common network status consensus, so that the evaluation in that interval is based on the view of a single directory; this explains the sharp decline in running relays in November 2007 which was not present in the views of other directories (which in turn would contain other such artifacts). The intermittent decrease of running relays in May 2008 can be explained by the Debian OpenSSL predictable random generator bug that led to blacklisting a certain number of relays by the directory authorities. The high variability of relays with `Stable` and `Guard` flags indicates a problem in the authorities assigning these flags that is currently under investigation, which is in parts the result of visualizing the problem as it is done here.

[3] Update August 27, 2012: the data is available at: `https://metrics.torproject.org/`

[4] Update August 27, 2012: these tools are no longer maintained, at least not for research purposes. The metrics website code at `https://gitweb.torproject.org/metrics-web.git` uses a database to aggregate relay statistics to answer standard questions, but it might be easier to develop custom database code to answer more research-oriented questions.

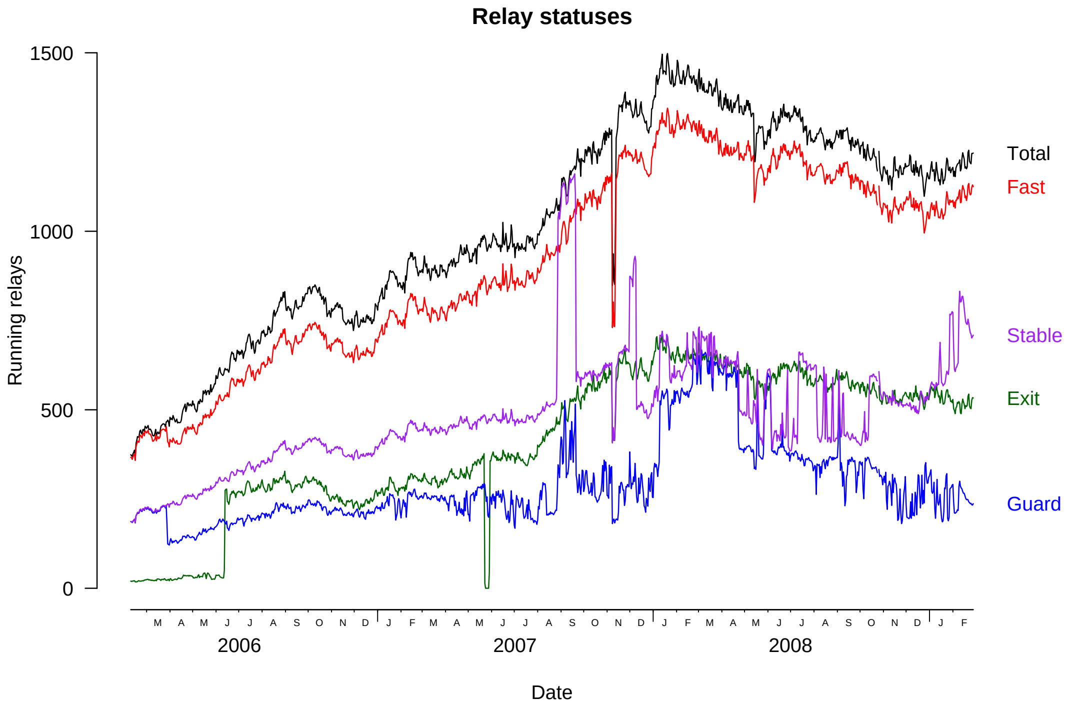

Figure 1: Average number of running relays per day with flags assigned by the directory authorities

This graph can also be useful to decide whether conditions to assign certain flags might require modification. For example, the average number of guard nodes has become rather low, given that these nodes carry one third of the total load of the network. One of the requirements for assigning the `Guard` flag to a relay is a weighted fractional uptime of at least 0.995, i.e., a relay was available for at least 99.5% of the time it is known to a directory authority. This number is simply a guess of the developers. Figure 2 shows the simulated effect of reducing the required weighted fractional uptime on the number of relays with the `Guard` flag. Simulations like this one based on real network data can be a useful tool for developers to estimate the consequences of design changes.

Relay versions. Relays include a platform string in their router descriptors containing the version of the Tor software and operating system. These strings can be evaluated to learn about the distribution of versions in the network as well as the update behavior of relay operators. Figure 3 visualizes the number of relays running different major versions of the Tor software. The vertical lines denote the points in time when a major version was declared to be the new stable version. The version life cycles can be subdivided into an alpha and release candidate phase (April 2006 to April 2007 for 0.1.2.x), a stable phase (April 2007 to July 2008), and an out-of-date phase (July 2008 until today). For all major versions there is an upper limit of approximately 200 relay operators running alpha or release candidate versions. There is no visible increase when versions are moved from alpha state to release candidate state (March 2, 2007 for 0.1.2.x, February 24, 2008 for 0.2.0.x). The stable phases for all versions show that it can take months until most relay operators switch from an out-of-date version to the stable

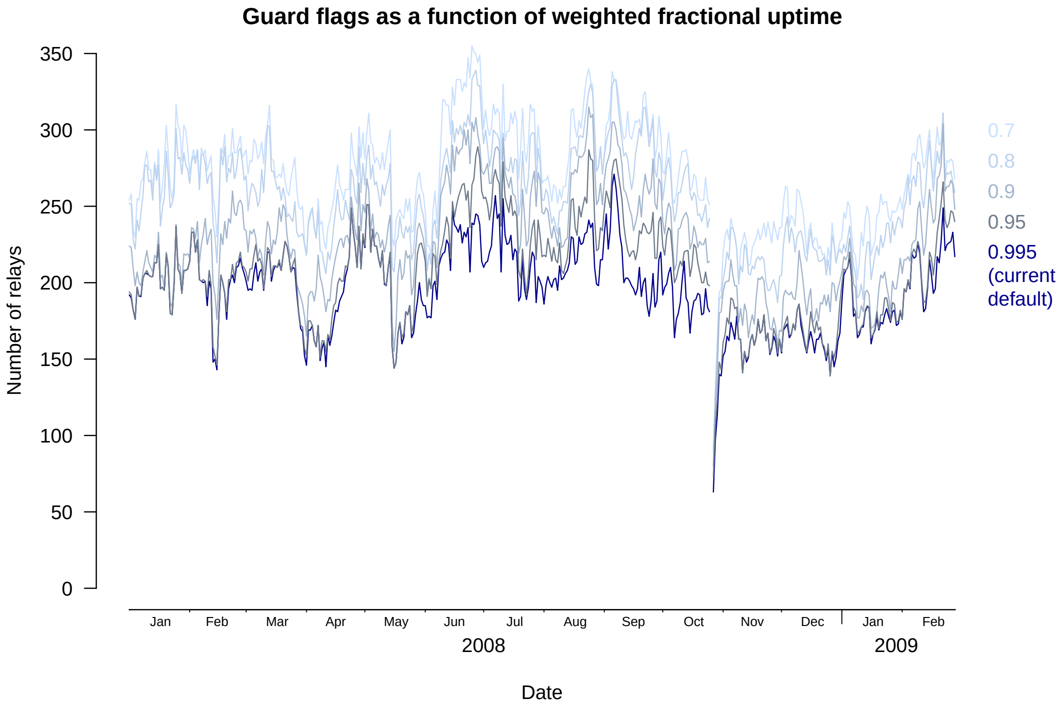

Figure 2: Simulation of the number of guard nodes when reducing the required weighted fractional uptime

version (April 2007 to around end of 2007 for 0.1.1.x). Accordingly, the out-of-date phases show that old versions are used even years after new stable versions are available (0.1.1.x still in use in 2009).

These results indicate that more efforts need to be taken to encourage relay operators to upgrade. It is desirable that relay operators upgrade to the stable versions as soon as possible or by no later than the end-of-life announcement. Approaches to make relay operators upgrade more quickly include helping popular operating system distributors to include up-to-date Tor versions or providing a semi-automatic updating mechanism to facilitate the upgrade process.[5]

Relays on dynamic IP addresses. The relays in the public Tor network are run by volunteers which are both individuals and organizations donating their bandwidth and processing power to the network. As a result, some relays are run on home computers that obtain a new dynamic IP address periodically or after a reconnect. After an IP address change, clients need to learn the new IP address of a relay in order to build circuits using it.

It is not immediately possible to determine for the analysis whether a relay uses a dynamic IP address or not. For this analysis we distinguish relays running on dynamic from static IP addresses from the total number of addresses that a relay has used throughout the analysis interval. Relays seen with only 1 or 2 addresses are considered to run on static IP addresses, with the rationale that they might have been moved to another location at most once while

[5] For more details about the secure updater for Tor, called Thandy, which is currently under development, see https://gitweb.torproject.org/thandy.git

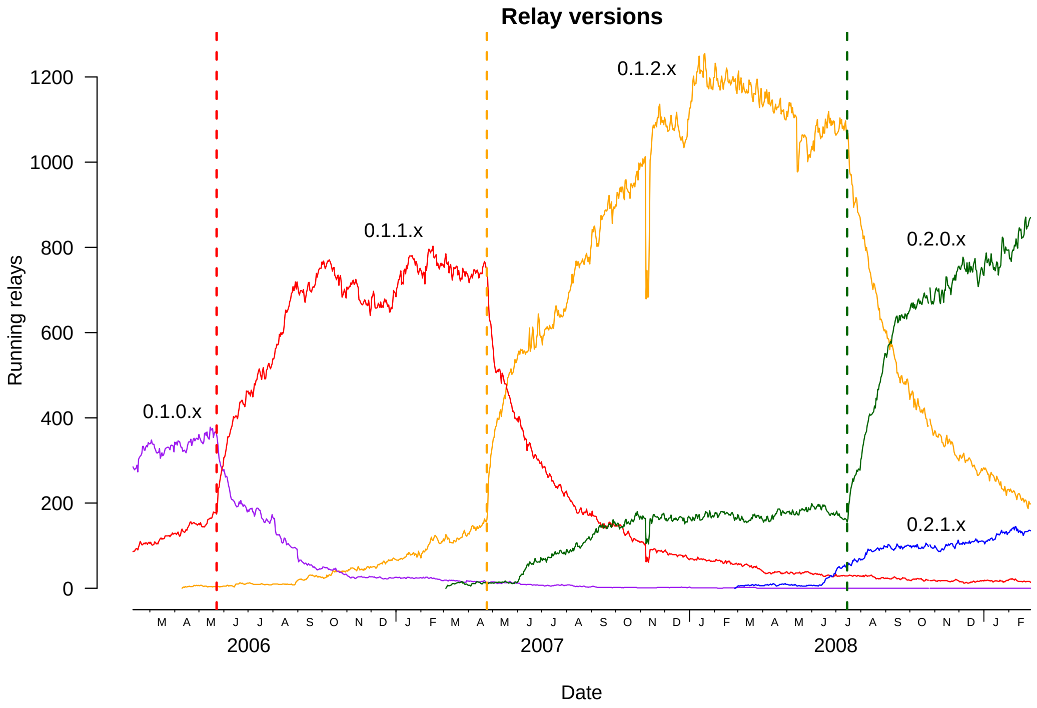

Figure 3: Number of relays running different major Tor software versions

keeping their identity. Relays that were seen with 3 or more IP addresses are considered to have dynamic IP addresses. More sophisticated ways to distinguish dynamic from static IP addresses, e.g., looking up addresses in a dynamic IP address database, have not been approached for this analysis.

Figure 4 shows the number of relays running on static and on dynamic IP addresses. The high number of relays on dynamic addresses indicates that efforts should be taken to make new relay addresses available to clients more quickly. Otherwise, a certain share of relays is unreachable for clients, leading to under-utilization of available bandwidth and a higher fraction of failed circuit build attempts. Interestingly, the decline of relays on dynamic IP addresses in 2008 has a similar pattern as the overall decrease of relays in that time.

Bandwidth capacity and usage. Relays report their observed bandwidth capacity and bandwidth usage to the directories. The bandwidth capacity is the maximum bandwidth as observed over any ten seconds in the past day. The idea is that this bandwidth peak constitutes the bandwidth that a relay is able to provide to its clients. The bandwidth capacity is used by clients during the path selection process to weight relays and obtain an overall load balancing in the network. Bandwidth usage is calculated as the total number of relayed bytes in 15-minutes intervals over the past day. This usage information is not considered by clients at all but is only made available for statistical purposes.

The graph in Figure 5 shows that roughly half of the available bandwidth capacity is used by clients. If the assumption is correct that relays can handle as much traffic as shown in the maximum 10-seconds interval over the past day, this indicates that the other half of the bandwidth remains unused. That would mean that better load balancing algorithms might make

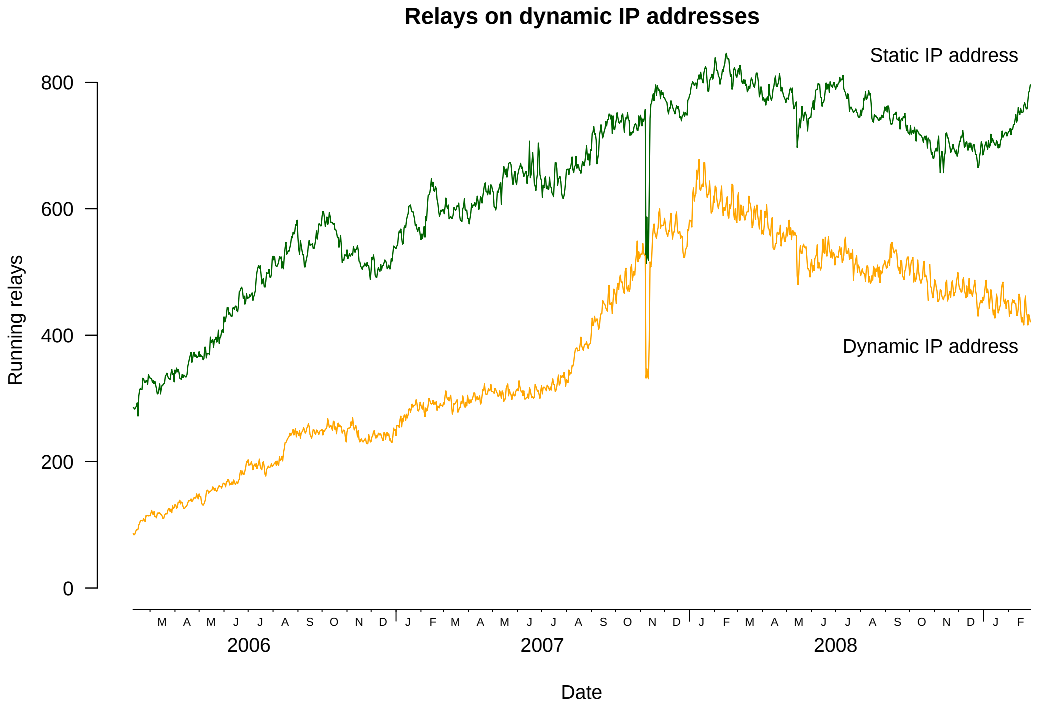

Figure 4: Number of relays presumably running on dynamic and static IP addresses

better use of the available bandwidth. On the contrary, the approach to measure bandwidth capacity as it is done today might be wrong and relays are not capable of handling that much traffic over longer periods. Further investigations are necessary to explain the discrepancy between provided and used bandwidth.

Relays by country. Finally, the archived directory data can be used to determine the locations of relays by using a GeoIP database. The distribution of relays to countries can give valuable insights into the willingness of people in certain countries to setup a relay. The trends can help detect possible problems that can be encountered by providing better support for relay operators in countries with decreasing numbers of relays.

Figure 6 shows the numbers of running relays in the top-5 contributing countries. The most visible trend is that the number of German relays suddenly stops growing in January 2008 and significantly shrinks over 2008. This trend might be the result of data retention laws and the uncertainty of relay operators whether running a relay is still legal or not. The graph indicates that the loss of German relays is responsible for the decline of relays in the network in 2008.

Figure 7 shows the top-5 contributing countries, this time by bandwidth usage. The pattern for German nodes in 2008 is similar to Figure 6. Another pattern is that French relays have suddenly seen less usage (which is a result of less advertised capacity) in July 2008. Finally, the Netherlands are the third largest provider of bandwidth, even though they did not show up in absolute numbers in Figure 6. Possible explanations for the sudden decrease of bandwidth provided by French relays and the high bandwidth-per-relay ratio of relays located in the Netherlands are single hosting companies who support (or have stopped supporting) the operation of relays. It might be beneficial to put more efforts into keeping good hosting

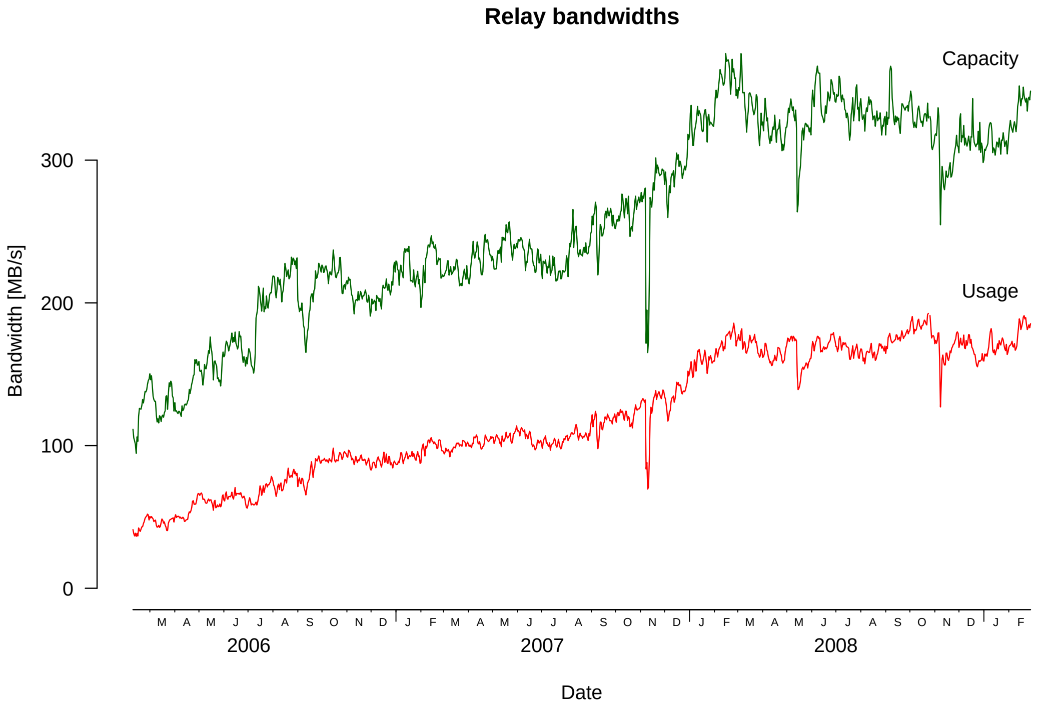

Figure 5: Total bandwidth capactiy and usage in the network

companies happy so that they continue to support individuals and organizations who wish to run relays.

4 Conclusion and Future Work

This paper presents an analysis of the Tor network by evaluating the existing directory information that is required for network operation anyway. The results show trends and reveal problems in the current network that need to be encountered, e.g., by lowering requirements for assigning certain flags, facilitating the upgrade process, improving support for dynamic IP addresses, possibly calculating bandwidth capacity more reliably, and clarifying legal issues for running relays in view of data retention laws. Statistical analysis of the network infrastructure can be a useful tool to detect problems and simulate or monitor the effect of changes.

The next step in analyzing the public Tor network would be to focus on performance and blocking-resistance metrics. The only data in the current analysis that is generated by users is bandwidth usage in terms of numbers of bytes in 15-minutes intervals. Future measurements could include more fine-grained network data in order to improve the Tor software and make it more useful. In contrast to the data presented here, more fine-grained network data would require collecting data that is not required for normal operation of the network. This raises concerns, both legally and regarding the security and anonymity of Tor users that need to be answered first, though. The work of McCoy et. al [6] has shown that additional measurements can be useful to learn more about the Tor network, but reinforces the necessity to settle possible liabilities before starting to collect more data.

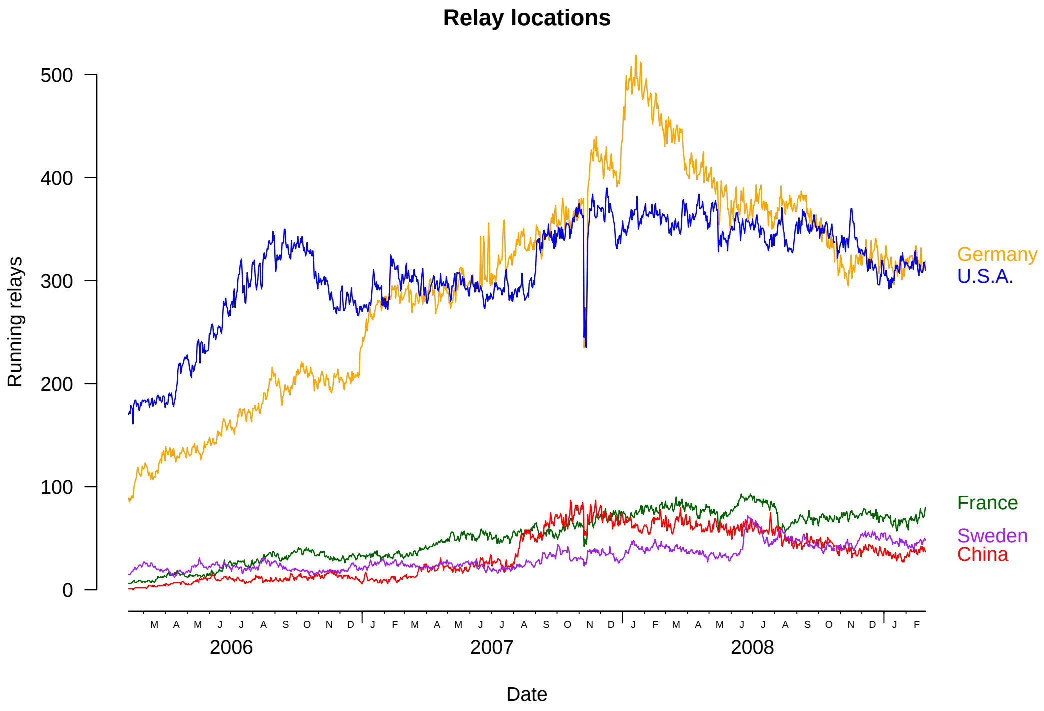

Figure 6: Number of relays in the top-5 contributing countries

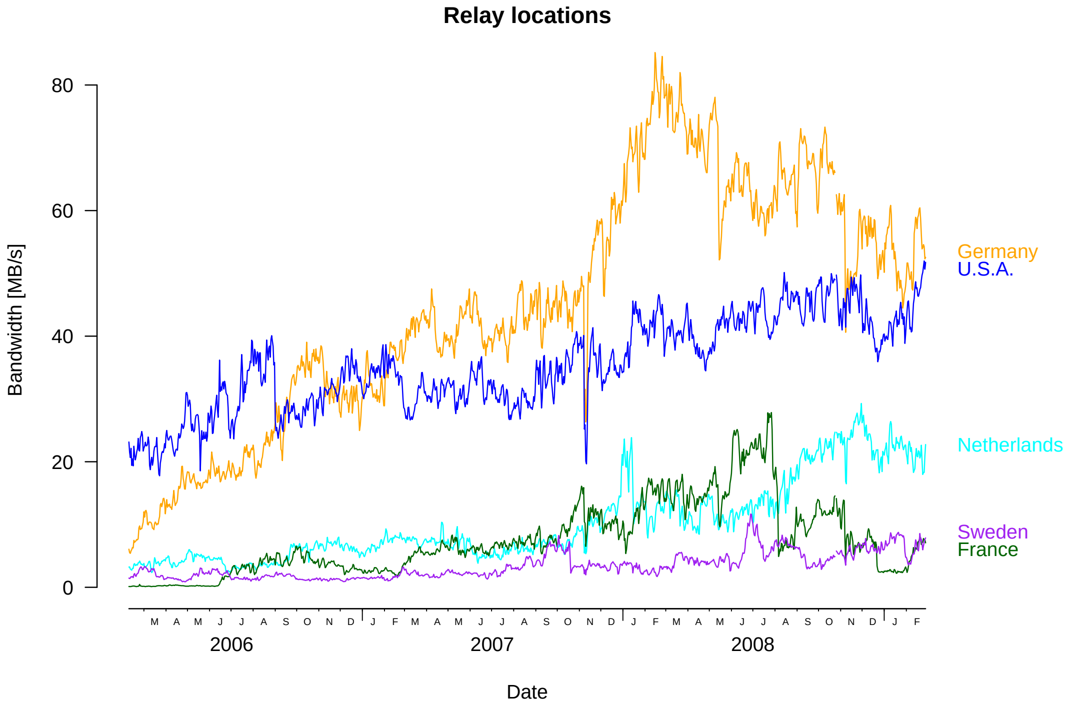

Figure 7: Bandwidth usage as observed by relays located in the top-5 contributing countries

References

[1] Kevin S. Bauer, Damon McCoy, Dirk Grunwald, Tadayoshi Kohno, and Douglas Sicker. Low-resource routing attacks against Tor. In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES 2007)*. ACM, October 2007.

[2] Roger Dingledine. Clients download consensus + microdescriptors. Tor Proposal 158, The Tor Project, January 2009. `https://gitweb.torproject.org/torspec.git/blob/HEAD:/proposals/158-microdescriptors.txt`.

[3] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium*, pages 303–320, August 2004.

[4] Jörg Lenhard, Karsten Loesing, and Guido Wirtz. Performance measurements of Tor hidden services in low-bandwidth access networks. In *Proceedings of the Seventh International Conference on Applied Cryptography and Network Security (ACNS 2009), Paris-Rocquencourt, France*, volume 5536 of *Lecture Notes in Computer Science*, pages 324–341. Springer, June 2009.

[5] Karsten Loesing. *Privacy-enhancing Technologies for Private Services*. PhD thesis, University of Bamberg, May 2009.

[6] Damon McCoy, Kevin Bauer, Dirk Grunwald, Tadayoshi Kohno, and Douglas Sicker. Shining light in dark places: Understanding the Tor network. In Nikita Borisov and Ian Goldberg, editors, *Proceedings of the Eighth Symposium on Privacy Enhancing Technologies (PETS 2008)*, volume 5134 of *Lecture Notes in Computer Science*, pages 63–76, Leuven, Belgium, July 2008. Springer.

[7] Martin Mulazzani. Personal communication, March 2009.

[8] Lasse Øverlier and Paul Syverson. Locating hidden servers. In *Proceedings of the Symposium on Security and Privacy (S&P 2006)*. IEEE Computer Society, May 2006.

[9] The Tor Project. *Tor directory protocol, version 3*, 2009. `https://gitweb.torproject.org/torspec.git/blob/HEAD:/dir-spec.txt`.