EAGER: Privacy-preserving measurements of the Tor network to improve performance and anonymity

Roger Dingledine, The Tor Project

1 Overview and Motivation

As the Tor network has grown since 2003 to almost 2000 volunteer relays, the anonymity that it can provide has grown too. With a user base now numbering in the hundreds of thousands, however, the performance of the network has suffered. We propose to measure the characteristics of Tor's network and usage, laying the foundation for evaluating its anonymity and improving performance.

Specifically, we will address three main components of this challenge. First, we will invent new algorithms for collecting Tor network load and usage data safely. We will need new metrics to assure we collect this data in a way that does not compromise users' privacy while providing useful research data. Second, we will collect and make available aggregated data about the live Tor network and its usage over time, and design and deploy new tools to manipulate and understand this data. Third, we will identify which measurements are necessary to support the wider performance and anonymity research questions, do the measurements, and feed the results into the anonymity community's ongoing research projects.

1.1 About Tor

Tor is a network of virtual tunnels that allows people and groups to improve their privacy and security on the Internet. As a platform, Tor enables software developers to create new communication tools with built-in privacy features. Tor provides the foundation for a range of applications that allow organizations and individuals to share information over public networks without compromising their privacy.

Individuals use Tor to keep websites from tracking them and their family members, or to connect to news sites or instant messaging services when these are blocked by their local Internet providers. Individuals also use Tor for socially sensitive communication: chat rooms and web forums for rape and abuse survivors, or people with illnesses.

Journalists use Tor to communicate more safely with whistleblowers and dissidents. Non-governmental organizations (NGOs) use Tor to allow their workers to connect to home websites while they are abroad, without broadcasting to everybody nearby that they are working with a possibly-sensitive organization.

Groups such as Indymedia recommend Tor for safeguarding their members' online privacy and security. Activist groups like the Electronic Frontier Foundation (EFF) recommend Tor as a mechanism for maintaining civil liberties online. Corporations use Tor as a safe way to conduct competitive analysis, and to protect sensitive procurement patterns from eavesdroppers.

A branch of the U.S. Navy uses Tor for open source intelligence gathering, and one of its teams used Tor while deployed in the Middle East. Law enforcement uses Tor for visiting or surveilling web sites without leaving government IP addresses in their web logs, and for security during sting operations.

The variety of people who use Tor contributes to its security [1]. Tor hides users among other users on the network, so a populous and diverse user base means better anonymity protections [6].

1.2 TOR MATTERS TO THE RESEARCH COMMUNITY

Just about every major security conference these days has a paper analyzing, attacking, or improving Tor. Examples just from 2008-2009 include Usenix Security [10, 25, 28], ACM CCS [9, 13, 20], PETS [2, 3, 19, 22, 26], and others [11, 27]. While ten years ago the field of anonymous communications was mostly theoretical, with researchers speculating that a given design should or shouldn't work, Tor now provides an actual deployed testbed. Tor has become the gold standard for anonymous communications research for three main reasons:

First, Tor's source code and specifications are open. Beyond its original design document [7], Tor provides a clear and published set of RFC-style specifications [5] describing exactly how it is built, why it made each design decision, and what security properties it aims to offer. The Tor developers conduct design discussion in the open, on public development mailing lists, and the public development proposal process [18] provides a clear path by which other researchers can participate.

Second, Tor provides open APIs and maintains a set of tools to help researchers and developers interact with the Tor software. The Tor software's "control port" [4] lets controller programs view and change configuration and status information, as well as influence path selection. We provide easy instructions for setting up separate private Tor networks for testing. This modularity makes Tor more accessible to researchers because they can run their own experiments using Tor without needing to modify the Tor program itself.

Third, real users rely on Tor. Every day hundreds of thousands of people connect to the Tor network and depend on it for the broad variety of security goals described in Section 1.1. In addition to its emphasis on research and design, The Tor Project has developed a reputation as a non-profit that fosters this community and puts its users first. This real-world relevance motivates researchers to help make sure Tor provides provably good security properties.

2 RESEARCH PLAN

Because of its consistent work on research, development, usability, and advocacy, Tor has already made a broad impact around the world. But the Tor network's operation in practice has produced emergent properties that require better understanding both for our developers and for other researchers. Through this project, we will position The Tor Project to provide the same level of openness for measurement tools and results as we already provide with respect to openness of source code and design decisions.

Along the way, we need to address hard theoretical questions about what data we must collect, and how to safely collect and aggregate it. In this section we outline two categories of specific data that we will analyze: directory and network data, and data about client and network performance.

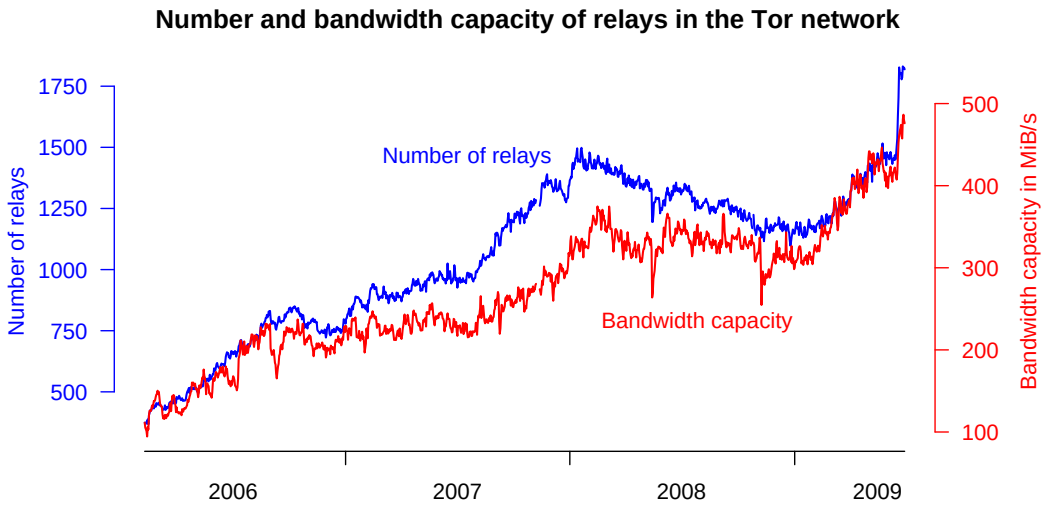

Figure 1: Number and bandwidth capacity of Tor relays from February 2006 to today

Research Activity 1: Directory and network data. The Tor network is an overlay network of volunteers running *Tor relays* that relay TCP streams for *Tor clients*. The client learns which relays it can use by fetching a signed network summary from the *directory authorities*. Specifically, each relay publishes a self-signed descriptor summarizing its address, bandwidth capacity, public keys, and other properties, and the authorities announce these descriptors along with opinions on whether each relay is considered reliable, fast, and so on.

The Tor Project has collected and archived these relay lists, descriptors, and directory authority opinions since mid-2004. As of June 2009, the directory archives have accumulated to a size of 23 GiB with an additional 1 GiB being added every month. We have recently started to analyze these archives to learn more statistics about the Tor network. As one example, Figure 1 shows the development of the number of relays and the bandwidth capacity of relays in the Tor network. The network grows to 1,500 relays through the end of 2007, drops throughout 2008 to 1,200 relays, and rises to 1,800 relays in June 2009. In the same time, the bandwidth capacity has grown more or less by the same percentage to a total capacity of almost 500 MiB/s.

Other results from evaluating directory information [15] show trends and reveal problems in the current Tor network that need to be addressed: clients use the directory authority opinions to make path selection decisions and to load balance better over the network, but the authorities are tuned poorly and are giving out opinions that lead to inefficient use of network resources; we need to work harder at getting relays to upgrade, since we still have many relays running old and possibly vulnerable Tor versions; we need better support for relays with dynamic IP addresses, to make these relays available to clients more quickly; we need to calculate bandwidth capacities more accurately and more reliably; and we need to better understand why we gain or lose relays (the drop in 2008 came from losing relays in Germany, perhaps from concern about their new data retention laws; whereas the spike in mid 2009 comes mostly from U.S. relays wanting to help activists in Iran).

The analysis of archived directory information also provides information about churn in the Tor network, giving us more accuracy when we simulate the effect of design changes [16].

One of the future challenges, besides analyzing the directory information more, is to make these data available to other researchers. Right now, researchers must collect directory information

themselves and then write their own parsers and evaluation tools. The Tor Project has recently made all the directory archives from 2004 to today publicly available at `http://archive.torproject.org/`. Alongside that project we will develop and make available tools to extract useful information from the directory archives. One valuable product of this process is that we will allow other researchers who are only indirectly working on metrics about the Tor network to make correct assumptions for their research.

Research Activity 2: Performance data. Connections over Tor exhibit high latency – and worse, high variance – which discourages the average user from routing traffic over Tor. In particular, the latency is higher than can be explained as the effect of redirecting traffic over three relays around the world plus the overhead of transferring bytes and performing cryptographic operations. Our initial investigations show that the problems come from congestion and queuing inside the relays.

Tor clients build *circuits* through three relays, and send traffic over the circuits in fixed-size *cells*, which are buffered at each relay until there is space in the relay's outgoing buffer. Relays use a token bucket for rate limiting, and refill the bucket once a second. Now that the Tor network is overloaded, these buffers often have more than one second's worth of cells in them, so the new behavior has turned into "send a burst of cells at the beginning of each second, and then send nothing until the next second." The effect of this design can be seen in Figure 2. The three lines show kernel density estimates of extension times to the first, second, and third hop of a circuit. All three lines exhibit an unusual accumulation of extension times at full seconds. For the first hop, there are only small bumps at 1 and 2 seconds, but for the second and third hop, these peaks become clearly visible at 2, 3, and 4 seconds. We suspect that the times that cells spend in queues until they are forwarded to the next hop are responsible for these peaks.

The same delays when extending existing circuits have turned out to be the number one reason why Tor *hidden services* are slow [12, 16, 17]. Tor hidden services allow users to offer low-latency services pseudonymously by interconnecting two Tor circuits on a common rendezvous point, thus protecting the locations of both client and server. While Tor circuits in normal operation can be built preemptively, so the delay for circuit construction doesn't impact the latency of page loads, hidden service circuits need to be extended on demand. Thus circuit extension delays introduce significant delays for hidden service page loads.

Similar to circuit creation and hidden services, delays due to congestion very likely affect transfer times of user data, too. We have started to analyze residence times of cells in circuit queues to obtain a general idea of how much time cells spend inside the network [14]. It also appears that this time is highly dependent on the loudness of a circuit, i.e. how many cells it has sent recently. It might be that the current round-robin scheduling algorithm that is applied to all circuit queues of a relay is not optimal: file-sharing traffic and other bulk transfer is taking much more than its fair share. Gathering further knowledge about the timing of cell processing is essential for designing and deploying improved scheduling algorithms.

Cell processing delays are only one reason for Tor's performance problems. We have identified five additional categories of reasons why the Tor network is slow, and sketched solutions for them [8]: some users put too much traffic onto the network relative to the amount they contribute, so we need to work on ways to limit the effects of those users and/or provide priority to the other users; further, an incentive system to prioritize users who contribute bandwidth might be worth considering [2, 23]; Tor's current path selection algorithms do not distribute load correctly over the network, so some relays are overloaded and others are under-utilized [22, 24]; clients should

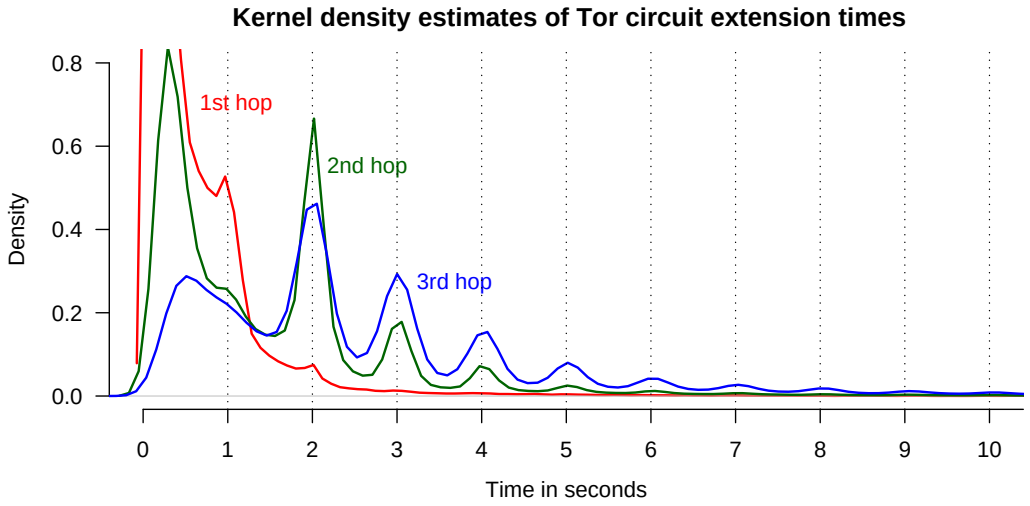

Figure 2: Kernel density estimates of extension times to first, second, and third hop in a Tor circuit

better handle variable latency and connection failures by dynamically adapting to changing network quality; and finally, low-bandwidth users spend too much bandwidth downloading directory information [12]. This project will find and analyze data to support (or refute) these theories.

3 Management Plan

The Principal Investigator for this two-year $300k project will be Roger Dingledine. Roger was the original developer for Tor, an original designer along with Nick Mathewson and Paul Syverson, and has been Project Leader for The Tor Project since its inception.

The bulk of the work on the project will be done by Roger Dingledine, Karsten Loesing, and Steven Murdoch, with help from other Tor staff. Karsten and Steven are the two core research staff working on Tor. Karsten wrote his doctoral thesis [16] studying Tor hidden services, including evaluating performance problems and designing improvements. Steven wrote his doctoral thesis [21] on anonymous communication, including a wide variety of attack papers on Tor.

4 Intellectual Merit and Broader Impact

A growing number of research groups are looking at Tor performance as a hot research topic, and we want to be ready to supply them with accurate and insightful assessments of usage in the deployed network. At the same time, The Tor Project has a variety of sponsors, such as Voice of America and other government agencies, who want to see performance improve. Once we understand why Tor is slow and how to fix it, other sponsors will fund us to deploy the improvements. Rigorous and well-grounded answers to the questions we raise in this proposal will be instrumental in making sure that our future work addresses the right problems.

References cited

[1] Alessandro Acquisti, Roger Dingledine, and Paul Syverson. On the economics of anonymity. In Rebecca N. Wright, editor, *Financial Cryptography*. Springer-Verlag, LNCS 2742, 2003.

[2] Elli Androulaki, Mariana Raykova, Shreyas Srivatsan, Angelos Stavrou, and Steven M. Bellovin. Par: Payment for anonymous routing. In Nikita Borisov and Ian Goldberg, editors, *Proceedings of the Eighth International Symposium on Privacy Enhancing Technologies (PETS 2008)*, pages 219–236, Leuven, Belgium, July 2008. Springer.

[3] George Danezis and Paul Syverson. Bridging and fingerprinting: Epistemic attacks on route selection. In Nikita Borisov and Ian Goldberg, editors, *Proceedings of the Eighth International Symposium on Privacy Enhancing Technologies (PETS 2008)*, pages 133–150, Leuven, Belgium, July 2008. Springer.

[4] Roger Dingledine and Nick Mathewson. Tor control protocol specification. `https://git.torproject.org/checkout/tor/master/doc/spec/control-spec.txt`.

[5] Roger Dingledine and Nick Mathewson. Tor protocol specifications. `https://git.torproject.org/checkout/tor/master/doc/spec/tor-spec.txt`.

[6] Roger Dingledine and Nick Mathewson. Anonymity loves company: Usability and the network effect. In *Proceedings of the Fifth Workshop on the Economics of Information Security (WEIS 2006)*, Cambridge, UK, June 2006. `http://freehaven.net/doc/wupss04/usability.pdf`.

[7] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium*, August 2004. `https://www.torproject.org/tor-design.pdf`.

[8] Roger Dingledine and Steven J. Murdoch. Performance improvements on Tor or, why Tor is slow and what we're going to do about it. Technical report, The Tor Project, March 2009. `https://www.torproject.org/press/presskit/2009-03-11-performance.pdf`.

[9] Matthew Edman and Paul Syverson. Astor: AS-awareness in Tor Path Selection. In *Proceedings of CCS 2009*, November 2009.

[10] Nathan Evans, Roger Dingledine, and Christian Grothoff. A practical congestion attack on tor using long paths. In *Proceedings of the 18th USENIX Security Symposium*, August 2009.

[11] Nicholas Hopper, Eugene Y. Vasserman, and Eric Chan-Tin. How much anonymity does network latency leak? *ACM Transactions on Information and System Security*, forthcoming 2009.

[12] Jörg Lenhard, Karsten Loesing, and Guido Wirtz. Performance measurements of Tor hidden services in low-bandwidth access networks. In *Proceedings of the 7th International Conference on Applied Cryptography and Network Security (ACNS 09), Paris-Rocquencourt, France*, June 2009.

[13] Zhen Ling, Junzhou Luo, Wei Yu, Xinwen Fu, Dong Xuan, and Weijia Jia. A new cell counter based attack against tor. In *Proceedings of CCS 2009*, November 2009.

[14] Karsten Loesing. Analysis of buffer sizes on a Tor middle node. Technical report, The Tor Project, May 2009. `https://git.torproject.org/checkout/metrics/master/report/performance/bufferstats-2009-05-25.pdf`.

[15] Karsten Loesing. Measuring the Tor network from public directory information. Technical report, 2nd Hot Topics in Privacy Enhancing Technologies (HotPETs 2009), Seattle, WA, USA, August 2009.

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

[17] Karsten Loesing, Werner Sandmann, Christian Wilms, and Guido Wirtz. Performance measurements and statistics of Tor hidden services. In *Proceedings of the 2008 International Symposium on Applications and the Internet (SAINT)*, Turku, Finland, July 2008.

[18] Nick Mathewson. The Tor proposal process. `https://git.torproject.org/checkout/tor/master/doc/spec/proposals/001-process.txt`.

[19] 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 International Symposium on Privacy Enhancing Technologies (PETS 2008)*, pages 63–76, Leuven, Belgium, July 2008. Springer.

[20] Jon McLachlan, Andrew Tran, Nicholas Hopper, and Yongdae Kim. Scalable onion routing with torsk. In *Proceedings of CCS 2009*, November 2009.

[21] Steven Murdoch. *Covert channel vulnerabilities in anonymity systems*. PhD thesis, University of Cambridge, December 2007.

[22] Steven J. Murdoch and Robert N. M. Watson. Metrics for security and performance in low-latency anonymity networks. In Nikita Borisov and Ian Goldberg, editors, *Proceedings of the Eighth International Symposium on Privacy Enhancing Technologies (PETS 2008)*, pages 115–132, Leuven, Belgium, July 2008. Springer.

[23] Johnny Ngan, Roger Dingledine, and Dan Wallach. Building incentives into Tor. Technical report, Rice University, November 2008.

[24] Mike Perry. TorFlow: Tor Network Analysis. Technical report, 2nd Hot Topics in Privacy Enhancing Technologies (HotPETs 2009), Seattle, WA, USA, August 2009.

[25] Joel Reardon and Ian Goldberg. Improving Tor using a TCP-over-DTLS Tunnel. In *Proceedings of the 18th USENIX Security Symposium*, August 2009.

[26] Micah Sherr, Matt Blaze, and Boon Thau Loo. Scalable link-based relay selection for anonymous routing. In Mikhail Atallah and Ian Goldberg, editors, *Proceedings of the Ninth International Symposium on Privacy Enhancing Technologies (PETS 2009)*, Seattle, August 2009. Springer.

[27] Robin Snader and Nikita Borisov. A tune-up for Tor: Improving security and performance in the Tor network. In *Proceedings of the Network and Distributed Security Symposium - NDSS '08*. Internet Society, February 2008.

[28] Sebastian Zander and Steven J. Murdoch. An improved clock-skew measurement technique for revealing hidden services. In *Proceedings of the 17th USENIX Security Symposium*, San Jose, CA, US, July 2008.