Poster: A Privacy-Preserving Protocol for Gathering Statistics About Tor Users

Ryan Henry
PhD Student
rhenry@cs.uwaterloo.ca

Tariq Elahi
PhD Student
mtelahi@cs.uwaterloo.ca

Ian Goldberg
Associate Professor
iang@cs.uwaterloo.ca

Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada N2L 3G1

(Poster Abstract)

Tor is a volunteer-operated anonymous communications network and associated free software package that facilitates low-latency, anonymous communications over the Internet. Current estimates place the number of daily Tor users at around 400,000 [1]. However, we stress that these are only very rough estimates; indeed, counting the number of Tor users without putting their anonymity at risk is a notoriously difficult problem [2]. Given this fact, it should be unsurprising that little is known about the characteristics of an "average" Tor user — if there *is* such a thing as an average Tor user. Several prior works (e.g., [3], [4]) have provided a partial glimpse into what such an average user might look like using information they obtained by instrumenting their own Tor relays (particularly, relays that act as ingress and egress points to the network). Alas, much of this data is now outdated and the consensus within the research community is that the privacy risks associated with such active measurements by Tor relays is unacceptable given current collection techniques. Even where the collected data is to be aggregated and anonymized before it is released, there are still legitimate concerns that the raw data might fall into adversarial hands. As such, the only up-to-date data available to researchers are those available through the Tor Metrics Portal [5], which publishes information about the Tor relays and the estimated numbers of Tor users (which can be broken down by country), but little more.

In this work, we present a novel protocol that enables privacy researchers to instrument their own Tor relays to log certain useful information about incoming and outgoing connections in a safe, responsible, and privacy-respectful way. Our approach combines 1) probabilistic data structures (e.g., hash tables or bloom filters), 2) threshold, partially homomorphic and semantically secure encryption, and 3) a secure shuffling protocol (similar to mental poker) to substantially eliminate the privacy risks inherent in logging usage statistics pertaining to individual Tor users. Our new protocol enjoys the following benefits:

1) **Deployability:** the protocol does not require any modifications to existing Tor clients nor to any other relay on the network.
2) **Data reliability:** the protocol relies on active measurements rather than on self-reporting of information by Tor users; thus, the validity of collected statistics does not rely on truthful reporting by Tor users or other untrusted relays.
3) **Confidentiality:** no potentially sensitive data is ever stored in plaintext; rather, all data is aggregated and stored using a partially homomorphic, semantically secure encryption scheme that conceals both the content and volume of data in the encrypted log file.
4) **Snapshot resistance:** an adversary that obtains two or more snapshots of the encrypted log file at different times learns nothing about what parts (if any) of the log file changed between the snapshots.
5) **(Threshold) forward secrecy:** appropriate use of threshold encryption ensures that an adversary that obtains the encrypted log file and all of the researchers' long term and short term secret keys cannot extract useful information from the log file (the adversary would also need to compromise the keys of some number of *trustees*, which could

include, for example, representatives for The Tor Project). Moreover, perfect forward secrecy ensures that, even in the unlikely event that such a breach of *all* trustees secret keys does occur, no useful information can be extracted from log files of past measurement periods.

6) **Safety:** All data is aggregated in encrypted form (i.e., prior to decryption); moreover, as long as a single trustee (or the researcher) is honest, the link between any particular Tor user and its associated data points in the log file are destroyed prior to decryption. (Of course, it is up to the researchers and trustees to collectively agree on what anonymized, aggregate statistics are acceptable to disclose.)
7) **Accountability:** extracting useful statistics from log files requires the consent and cooperation of all trustees; thus, if there are unanticipated revelations or events that could cause hitherto innocuous statistics to become potentially dangerous to disclose, the trustees can simply refuse to assist in the decryption process.
8) **Efficiency:** the protocol is efficient enough to collect data in real time, even on high capacity, heavily-utilized relays. (Decryption is a more expensive — though manageable — offline process.)

We fully expect that our new approach will make a significant positive impact on Tor by enabling Tor researchers to investigate many new lines of inquiry that were previously dangerous, and therefore taboo, to study on the live Tor network.

References

[1] "Tor Metrics Portal: Users," The Tor Project, April 2012, [Online; accessed April 6, 2012] https://metrics.torproject.org/users.

[2] S. Hahn and K. Loesing, "Privacy-preserving Ways to Estimate the Number of Tor Users," The Tor Project, Technical report, November 2010, https://metrics.torproject.org/papers/countingusers-2010-11-30.pdf.

[3] S. Le-Blond, P. Manils, C. Abdelberi, M. A. Kâafar, C. Castelluccia, A. Legout, and W. Dabbous, "One Bad Apple Spoils the Bunch: Exploiting P2P Applications to Trace and Profile Tor Users," *CoRR*, vol. abs/1103.1518, 2011.

[4] D. McCoy, K. S. Bauer, D. Grunwald, T. Kohno, and D. C. Sicker, "Shining Light in Dark Places: Understanding the Tor Network," in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, N. Borisov and I. Goldberg, Eds., vol. 5134. Springer, 2008, pp. 63–76.

[5] "Tor Metrics Portal," The Tor Project, April 2012, [Online] https://metrics.torproject.org/.