Privacy-preserving Anomaly Detection in Tor

Tariq Elahi
ESAT-COSIC
KU Leuven
tariq@kuleuven.be

Ryan Henry
School of Informatics and Computing
Indiana University Bloomington
henry@indiana.edu

Abstract

This extended abstract presents our vision of *PrivEy*, a distributed data collection and anomaly detection framework for the Tor network. *PrivEy* builds on the general framework of *PrivEx* (CCS 2014), a system for privately collecting statistics about traffic egressing the Tor network; however, *PrivEy* extends *PrivEx* in several important respects: (i) it supports the collection of a wider array of data from a wider array of vantage points within the Tor network, and (ii) beyond merely producing differentially private summary statistics about the collected data, it can also use those data to continuously train ensemble classifiers with which to recognize anomalous patterns indicative of ongoing attacks against the Tor network and its users.

1 Introduction and motivation

Privacy is a fundamental human right [1] whose value is recognized and codified in laws around the world, and anonymity is an essential tool for preserving the privacy of individuals in today's increasingly online world. Anonymous communications networks—the most notable example of which being Tor [2]—are therefore essential tools for protecting human rights. Indeed, millions of users worldwide depend upon the anonymity that Tor provides each and every day [6]. These users span the spectrum from activists, whistleblowers, and citizens of oppressed countries who use Tor to overcome censorship and to safely expose abuse and corruption, through to journalists and members of law enforcement who use Tor to protect their sources and preserve the integrity of their investigations, all the way to regular folks who use Tor to keep the intimate details of their everyday activities free from prying eyes.

In many instances, these users rely on Tor for their physical, emotional, and financial well-being; as such, the Tor community considers it a moral imperative to act judiciously and cautiously when it comes to collecting potentially sensitive data about the Tor network and its users. The current zeitgeist holds that collecting Tor data is fraught with risks to user privacy, however vaguely defined, and should therefore be avoided whenever possible. Alas, a recent spate of high-profile incidents [4, 5, 8] starkly illustrate the fact that indiscriminately eschewing data collection can be a double-edged sword: data about the goings-on of the Tor network might contain clues to assist in the early detection and mitigation of attacks that would otherwise jeopardize the safety of the very users the Tor community's nigh-prohibition on data collection seeks to protect.

Despite ongoing debates,[1] the inherent risks posed by data collection on the Tor network are very poorly understood and, consequently, there exists disparagingly little consensus among the Tor community regarding which *kinds* (and *granularity*) of data are "safe" to collect. The framework we propose herein is guided by our own views on this matter, which hold that Tor data should be considered "safe" to collect and disclose if (and perhaps only if) any miscreants who might

[1] Indeed, the tension between collection and non-collection of Tor data is so acute that The Tor Project's board of directors recently formed the *Tor Research Safety Board* (https://research.torproject.org/safetyboard.html) as an informational and advisory body aiming to help researchers collect vital data about the Tor network while avoiding unnecessary risks to the privacy of its users.

Private Multi-Party Machine Learning – NIPS 2016 Workshop, Barcelona, Spain.

conceivably leverage said data to compromise the privacy of Tor users could just as easily collect the same data (or, at least, some equivalently damning data) by launching a "low cost" and undetectable attack. This is a somewhat more *permissive* standard of safety than the status quo, as it explicitly allows for some (albeit limited) amount of "harm" to result from the data collection—indeed, perhaps even harm which could not occur in a parallel universe whose denizens are all honest-but-curious—and yet (once one fixes a suitable definition of "low cost") it is also a more *objective* standard that, we contend, is not likely to engender real privacy violations in *this* universe, where typical attackers are comparatively well funded and, though certainly curious, far from honest. Thus, while our ideas undeniably push the envelope, so to speak, of acceptable Tor data collection, we nonetheless expect (or, at least, hope) that our proposal will be decidedly non-controversial within the Tor community.

Outline. We begin in Section 2.1 with a bird's-eye view of our proposed architecture, after which we flesh out a detailed threat model in Section 2.2, before returning to the architecture in Section 2.3 to expand on and defend some of its key properties. We conclude in Section 3 with a discussion of the key challenges in our proposal. Throughout this discussion, we assume that the reader has a basic familiarity with Tor's architecture.

2 The proposal

2.1 Framework

We envision a distributed data collection and anomaly detection apparatus for Tor, with the eventual goal of being able to collect and act upon data about large volumes of traffic flowing into, through, and out of the Tor network in nearly real-time. This data will be gathered by a designated subset of Tor relays (called *data collectors*, or *DCs*) that have each been instrumented with one or more special *data collection modules*. Each data collection module will contain the logic required to capture, store, and process data required to realize some particular functionality, such as detecting anomalous network behaviour or measuring the prevalence of specific types of Tor usage. The DCs will then periodically run secure multiparty computation (MPC) protocols to either aggregate and publish summaries of the data (in differentially private form), or to train and evaluate the predictions of machine learning (ML) ensemble classifiers in order to detect anomalous patterns indicative of problems with or ongoing attacks against the Tor network and its users.

We tentatively call our proposed framework *PrivEy* to highlight its status as the "natural successor" to *PrivEx*, a privacy-preserving data collection system recently proposed by Elahi, Danezis, and Goldberg [3] with the related-yet-much-more-modest goal of quantifying how much of the traffic flowing out of the Tor network has been (or, at least, appears to have been) routed through Tor as a way to bypass state-level censorship. Although we will lay much of the groundwork for *PrivEy* in what follows, our proposal is still very much in the "working" phase and the feedback that we hope this extended abstract will generate may well inspire important differences between the details of our exposition herein and the final form our system ultimately takes. Specifically, while the first author's experience designing and implementing *PrivEx* provides us with a solid understanding of how to safely collect and store Tor data, we invite discussions with experts from the machine learning (ML) community on how best to realize our vision of running ML algorithms to classify the collected data and to provide nearly real-time anomaly detection for the Tor network.

We are particularly interested in leveraging ML techniques to understand network activity, such as discovering the types of traffic (i.e., protocols) that traverse the network, classifying traffic by usage patterns (bulk transfers vs. interactive sessions), and detecting anomalous circuit construction/destruction behaviour. We stress that the problem setting we target necessitates very strong privacy guarantees, and we are therefore willing to tradeoff relatively large amounts of precision and accuracy in exchange for the strongest possible privacy protections.

Our working model involves a (conceptual) data aggregator, which will be realized in practice using a secure MPC protocol run among the DCs (and possibly some other trustees). Each DC will collect data on a continuous basis, and periodically the DCs will come together either to produce differentially private summaries of the combined data, or to update a "global" ML model using their respective datasets. Indeed, the differentially private data summaries may be used to inform the "global" ML model as an alternative realization of the aggregation process.

2.2 Threat model

Our threat model distinguishes between two distinct classes of attackers (which we call *subversive attackers* and *inference attackers*) based on the attackers' agendas. Consistent with Tor's current

threat model, we constrain both types of attackers to be non-global; that is, we assume that the attackers are localized to some portion of the Tor network, so that they can only observe or tamper with communications that pass through that portion. We defend this restriction to non-global attackers by noting that any data one can hope to gather with *PrivEy* would be rather trivial for a global attacker to learn through direct observation.

Note that both types of attacker can leverage auxiliary information and resources that are not available to honest DCs. For example, in addition to wielding complete control over some coalition of Tor relays (perhaps including some DCs), the attacker might control a website that some users access via Tor, or it might reside at an ISP that can observe and tamper with connections originating from its own subscribers; likewise, the attacker might be in a position to perform numerous queries against the ML-trained classifier model in an attempt to infer *sensitive* information about the training set used [7]. Due to the availability of such external information and influence, it will generally not be feasible to guarantee that *PrivEy* reveals zero (or even some reasonably small ε-level of) potentially sensitive information to every conceivable attacker; however, in keeping with our proposed standard of "safety", we will seek to design mechanisms for which we can prove that, whatever sensitive information *PrivEy* does reveal to a given attacker, that same attacker could have easily inferred as much through a low-cost and undetectable attack. (In particular, a motivated attacker should learn essentially the same quantity of sensitive information irrespective of the existence or non-existence of *PrivEy*.)

We now briefly describe the motivations and potential tactics of the two attacker types.

Subversive attacker. A subversive attacker is akin to the sort of attacker typically considered in a cryptographic protocol: it is intent on violating the security guarantees offered by the cryptography being used to protect and compute with raw, sensitive data about Tor users. The subversive attacker controls a subset of the DCs and is not constrained to follow any portion of the protocol correctly. We typically assume that the subversive attacker behaves *covertly*, meaning that it will only deviate from the prescribed protocol in ways that are likely to go unnoticed by non-colluding DCs or external observers.

We assume that the subversive attacker is capable of wresting control of previously honest DCs—either via relay compromise or through legal coercion—in a bid to learn whatever data those DCs have collected to date. While it is impossible to prevent a compromised relay from storing information that it should not store (or, at least, that it should not store in the clear), we do seek to frustrate attempts by a subversive attacker to expose data that was collected by a DC prior to its being corrupted.

Rather than attempting to learn about sensitive raw data, a subversive attacker may try to simply poison the inputs to/outputs from *PrivEy*, for example to prevent an anomaly detection module from alerting defenders about an ongoing attack.

Inference attacker. The inference attacker is akin to the sort of attacker typically considered in the settings of differential privacy and data anonymization: it is intent on leveraging auxiliary information to draw inferences beyond what *PrivEy* is intended to allow. The inference attacker may control resources both internal and external to the network, and can combine data from all sources within its purview.

In an attempt to subvert the aggregation and ML protocols, and thereby reveal information that would otherwise remain private, an inference attacker who controls external resources (such as websites or Internet infrastructure) may manipulate the operation of these resources to introduce biases and correlations in data observed by honest DCs; likewise, an inference attacker that controls one or more DCs may selectively withhold some legitimate data and/or inject bogus data into the datasets it holds.

Note that a single attacker can don both a subversive hat and an inference hat at the same time.

2.3 Solution sketch

Much like its predecessor *PrivEx*, our proposed *PrivEy* framework runs atop and interfaces with the existing Tor infrastructure in a way that does not require any modifications to Tor clients nor to existing relays that are not part of the *PrivEy* data collection apparatus. As previously explained, a subset of the relays will be designated as DCs, and each DC will be furnished with a portfolio of data collection modules that manage the safe collection, storage, and eventual use of data about certain aspects of the network activity that DC is naturally able to (*passively*) observe as a byproduct of its position within the network.

A given DC may run several data collection modules at once; however, it is not required (nor necessarily desired) that every DC runs every data collection module. For instance, certain data collection modules may be available only to those DCs that have the "guard" flag in the network consensus, while others may be available only to those DCs that have the "exit" flag. Moreover, depending on the nature of the data that a given data collection module seeks to gather, it might be necessary to prohibit any given DC from running two specific modules concurrently, lest correlations in the data recorded by the two modules reveal sensitive information about Tor users.

Despite comprising only passively observable data, it is imperative that the DCs treat the datasets they compile as being *extremely* sensitive. In order to mitigate against the subtle privacy harms that can arise when data available to one relay is combined with those available to another, Tor relays currently maintain essentially no logs about network activity. As the *PrivEy* collection apparatus will require DCs to begin logging certain information about network activity, we must take special care to ensure that the raw data held by the DCs cannot facilitate such privacy harms. To this end, we propose the following precautions: (i) the DCs should only collect data at the coarsest level of granularity for which it is still possible to measure the desired effect (this may involve hashing, generalizing, bucketizing, and/or aggregating data as it is collected), and (ii) the datasets should be encrypted at all times, and the encryption should leverage threshold techniques so that even the DC holding a given dataset is unable to decrypt it. Even with these precautions in place, some kinds of data may simply be too sensitive to collect. Which data should and should not collected by *PrivEy* data collection modules is a decision that the Tor community should collectively make through open discussions, perhaps facilitated by the Tor Research Safety Board; in any case, discussing the intricacies of collecting specific data goes well beyond the scope of this extended abstract.

Periodically, all of the DCs running a particular data collection module will run a secure MPC protocol on the datasets they collected through that module. The frequency and nature of this MPC will vary depending on the characteristics of the data collected and how that data will be used. There are two basic use cases for the data:

1. The DCs can aggregate their datasets and then compute specific, differentially private statistics over the combined data. These statistics might, for example, help users understand the "health" of the network so as to gauge the level of anonymity they can expect over time; they might help developers identify bottlenecks or software bugs; or they might help researchers and funders understand how the Tor network is and is not being used.
2. Each DC can use its own (individually collected) data to train a classifier using an appropriate ML algorithm optimized to detect specific types of malicious behaviour. The localized models from the DCs can then be aggregated into a single, global model that can detect —in near-real time—anomalous network behaviour and alert network and node operators to take mitigatory actions.

3 Obstacles and opportunities

Our envisioned *PrivEy* system pushes the boundaries of what is (currently known to be) possible using ML, secure MPC, and differential privacy building blocks. The critical limitations on widespread deployment seem, at this time, to be concerned with the practicalities of fielding such a system: for *PrivEy* to see the light of day, it will be vitally important to strike a fine balance between the strong security and privacy requirements of the Tor community, on the one hand, and the utility and efficiency requirements needed to make such a system practical, on the other hand. Indeed, given that Tor is powered by a conglomeration of volunteer-operated relays, it is necessary to ensure that the overhead imposed by *PrivEy* is not especially burdensome to the DCs from a computation, storage, or communication perspective; however, we still need to ensure that *PrivEy* can effectively detect anomalies and respond to them in nearly real time, all the while maintaining the Tor communities' stringent privacy requirements.

This is all easier said than done, and will necessitate a considerable amount of cryptographic engineering and basic research in the intersection between ML and differential privacy. For instance, it is currently unclear (to us) at what points one should apply DP—to the training set, to the ML models that form an ensemble, to the ensemble model itself, or to some combination of three? Different choices here will have different implications for not only for the kinds of predictions the ML model can make, but also for the computational efficiency of the DP mechanism. Yet meeting these challenges will be a productive step towards a healthier and more robust Tor networks that, going forward, can continue to defend the privacy and basic human rights of its users.

References

[1] UN General Assembly. Universal declaration of human rights (December 1948). Adopted and proclaimed by General Assembly resolution 217 A (III). http://www.refworld.org/docid/3ae6b3712c.html.

[2] Roger Dingledine, Nick Mathewson, and Paul F. Syverson. Tor: The second-generation onion router. In *Proceedings of USENIX Security 2004*, San Diego, CA, USA (August 2004).

[3] Tariq Elahi, George Danezis, and Ian Goldberg. PrivEx: Private collection of traffic statistics for anonymous communication networks. In *Proceedings of CCS 2014*, pages 1068–1079, Scottsdale, AZ, USA (November 2014).

[4] The Tor Project. Did the FBI pay a university to attack Tor users? https://blog.torproject.org/blog/did-fbi-pay-university-attack-tor-users. [online; accessed 2016-09-22].

[5] The Tor Project. How to handle millions of new tor clients. https://blog.torproject.org/blog/how-to-handle-millions-new-tor-clients. [online; accessed 2016-09-22].

[6] The Tor Project. Tor Metrics — Direct users by country. https://metrics.torproject.org/userstats-relay-country.html. [online; accessed 2016-09-25].

[7] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *arXiv:CoRR*, abs/1610.05820 (October 2016).

[8] Philipp Winter, Richard Köwer, Martin Mulazzani, Markus Huber, Sebastian Schrittwieser, Stefan Lindskog, and Edgar R. Weippl. Spoiled onions: Exposing malicious Tor exit relays. In *Proceedings of PETS 2014*, volume 8555 of *LNCS*, pages 304–331, Amsterdam, Netherlands (July 2014).