Privacy-Preserving Longevity Study of Hidden Services

Motivation

Very little is known about the lifespan of hidden services. Such knowledge provides manifold benefits, such as the detection of malicious and benign domains. For example, it allows investigation of the maliciousness of domains, based on their lifespan. Short-lived hidden services could indicate not to be legitimate domains, as compared to long-lived domains. Moreover, all meta-data leakage around a new hidden service should be minimal. Ideally, a new hidden service behaves exactly like (the majority of) all hidden services, i.e., has typical lifetime. Otherwise, this hidden service will strike out of an anonymity set. Finally, such knowledge provides insights into the performance and resource allocation requirements of privacy infrastructures. In conclusion, acquiring additional knowledge about hidden service usage will be a benefit for regular Tor users.

However, to avoid disrupting Tor and its security and privacy services, any study must be carried out in a privacy preserving manner. The distributed nature of Tor and hidden services makes such study non-trivial and introduces challenges that need to be addressed. Along the same lines as the plethora of previous studies on Tor, we carefully implement privacy protection as follows.

Data Collection

Roughly speaking, we count the number of times a hidden service descriptor has been uploaded to an HSDir in an encrypted fashion. This count will later serve as an indicator of lifespan. We have analyzed that by running a certain number of HSDirs, we will be able to estimate the lifespan of hidden services with a certain error probability. To maintain the privacy services of Tor, we only seek an aggregate PDF/CDF of the lifetime of the hidden services with the following features:

- We use an additively homomorphic public-key encryption scheme with secret sharing. Here, n HSDir servers jointly compute a private/public key pair. The public key is known by everybody, but no coalition of $n-1$ HSDir servers can restore the private key.
- Each HSDir server features a hashtable, initialized with encryptions of zero. Storing a new onion address a on the HSDir triggers an action where the ciphertext at hashtable index $H(a)$ is modified by an homomorphic addition with an encryption of “1”.
- At the end of the measurement phase, three rounds take place. First, all HSDir servers homomorphically add their hashtables to aggregate counters. Second, in a daisy-chain manner, each HSDir server permutes the aggregated hashtable. Third, using secure multi-party computation, we realize an ideal functionality that decrypts the aggregated counters in random order and publishes them to us. The only information we learn is the sum of randomized counters, i.e., a histogram. From this, we derive the lifetime PDF/CDF.

Please note that

- Obviously, the raw (original) onion addresses are discarded.

- No information about a specific onion is leaked. The only information eventually revealed is a (permuted) histogram of lifetimes.
- An adversary compromising $n-1$ HSDir servers cannot infer anything about counters or onion addresses.
- Once data is encrypted even the HSDir relay holding cannot decrypt.
- Only when all the parties are involved, we will be able to decrypt. If one party is not participating the data remains irretrievable.

Implementation and Computation

We stress that the above approach to perform privacy-preserving “statistics” is not new. There has been a flurry of publications on the subject, and our techniques for counting and statistics resemble those by, e.g., [1], i.e., members of Tor’s research safety board. Computations are done by three different research entities, and no party has access to any clear data at any time. Furthermore, only when all parties participate in the computations, they can provide aggregate statistics. From the operational security we safeguard and insure standard privacy properties of Tor. No information is written to persistent storage, and data only briefly resides in RAM. Again, we run relays controlled by three different entities over two continents with different jurisdictions, so avoid coerced disclosure of data.

We are aware that the upcoming changes on next generation hidden services will jeopardize such research efforts.

[1] Tariq Elahi , George Danezis , Ian Goldberg, PrivEx: Private Collection of Traffic Statistics for Anonymous Communication Networks, Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014.