Poster: Practical Website Fingerprinting on Tor

Tao Wang
PhD Candidate
Cheriton School of Computer Science
University of Waterloo

Ian Goldberg
Associate Professor
Cheriton School of Computer Science
University of Waterloo

I. Background

Web users of Tor protect their anonymity by communicating to web servers through proxy relays using layered encryption. Attackers and eavesdroppers situated at the user's network or the first relay's network cannot read the contents of these messages, including the destination, because of encryption. However, it has been known [1], [2], [3] that Tor traffic may be susceptible to website fingerprinting, wherein the attacker guesses the true destination by leveraging packet sequence patterns, such as size, order, and timing.

In the above works, the attacker is given a packet sequence which corresponds to a single web page load, and asked which web page this sequence came from. This is an unfair assumption that benefits the attacker, as a practical attacker cannot trivially identify the start and end of each packet sequence. Real users may load partial pages, continuous pages, several pages at once, and they may not be loading pages at all. This implies that previous website fingerprinting works cannot be practically implemented until this detection problem is solved.

II. Impact

Attackers using website fingerprinting pose a significant threat to the privacy expected from Tor. A number of parties have an interest in conducting such an attack. Totalitarian government agencies, nervous about being unable to conduct surveillance on an anonymity network, may seek to do so with website fingerprinting. They may also seek to observe and classify anonymous traffic in order to justify a ban by demonstrating the potential prevalence of illegal activity. It could be possible to even block specific sites from Tor. On the other hand, attackers may attempt to track, identify, and observe users who expect privacy. We need to fully understand the potential impact of website fingerprinting. As it stands, anonymity network users, for instance, may be misled into thinking that their privacy is guaranteed and no one can determine their destination server, when it is all but certain when faced with website fingerprinting.

III. Our Contributions

In this poster, we will present a three-stage strategy to solve the detection problem of website fingerprinting. This problem has not been tackled or addressed in previous work, and our strategy is novel. The attacker starts with a sequence of packets generated by a simulated client, which contains multiple pages and possibly other activities such as file downloading.

In the first stage, we attempt to remove noise. This includes Tor noise such as SENDME packets and circuit construction packets, and also other activities such as file downloading and web chatting. To remove such noise, we use machine learning techniques with pattern matching and packet counting. SENDME packets are sent at regular intervals and therefore may be removed by packet counting. File downloading and web chatting will have significantly different packet patterns from normal web page loading, such as a different ratio and speed of incoming and outgoing packets, so we will show how these activities can be detected and removed to reduce noise.

Second, we look at rough metrics to find the start points of each packet sequence. These metrics include the ratio of incoming and outgoing packets, packet transmission rate, and others. The start of a page load has more outgoing packets and a higher packet transmission rate as the number of connections allowed by the browser is saturated. Using these metrics will help us determine where packet sequences begin, but with low precision.

Finally, we perform pattern matching to find the exact start points of each packet sequence. When web pages are loaded, the browser starts with a server connection request. Then, when incoming packets indicate it is accepted, the server sends an HTTP GET request for the main page, which is then followed by main page content. With Tor, users may send a HTTP GET request optimistically before seeing the server connection accept. We will show that such a sequence is easy to identify and can be used to determine the start point of the packet sequence.

A number of different strategies can be used for each stage, and we will show how effective such strategies are by performing them on data by simulating realistic clients on the live Tor network.

References

[1] X. Cai, X. Zhang, B. Joshi, and R. Johnson. Touching from a Distance: Website Fingerprinting Attacks and Defenses. In *Proceedings of the 19th ACM Conference on Computer and Communications Security*, pages 605–616, 2012.

[2] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel. Website Fingerprinting in Onion Routing Based Anonymization Networks. In *Proceedings of the 10th ACM Workshop on Privacy in the Electronic Society*, pages 103–114, 2011.

[3] T. Wang and I. Goldberg. Improved Website Fingerprinting on Tor. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, 2013.