Computer Communications 000 (2016) 1–9

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

A novel passive website fingerprinting attack on tor using fast fourier transform

Hojjat Jahani, Saeed Jalili*

Computer Engineering Department, Electrical and Computer Engineering Faculty, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Article history:
Received 31 October 2015
Revised 24 May 2016
Accepted 30 May 2016
Available online xxx

Keywords:
Anonymity network
Tor
Privacy
Website fingerprinting
Passive attack

ABSTRACT

One of the main applications of low latency anonymity networks, such as Tor, is to protect data and users' privacy from interception over the Internet. This paper presents a novel passive website fingerprinting attack to defeat Tor's anonymity mechanism with a significant improvement in comparison with similar methods. Unlike the state-of-the-art approaches, the proposed method does not need to change the sequence of IP packets exchanged between users and the first relay in the network. We introduce a new method based on the Fast Fourier Transform to calculate the similarity distance of instances from traffic patterns. In this way, the time complexity of feature extraction is reduced by a factor of 400 during the classification process. Considering a closed-world scenario, our method easily spots on target websites with a minimum success rate of 95%. As yet another notable payoff, the accuracy keeps up more robustly while the polarity of target website grows gradually.

1. Introduction

The fast growing trend of the Internet has introduced various approaches in data communications security (such as confidentiality and unlink-ability) and anonymity. SSL and VPN have made it possible to realize data security, while JAP [1] and Tor [2] technologies have combined both features of data security and anonymity together. Tor is one of the most famous anonymity technologies based on onion router, which is used by many people to benefit from such aforementioned features. Although Tor preserves the anonymity of users (including their ids and other info), but it is unable to prevent the information leakage of traffic flow, such as packet size, their frequency, direction of packets, and etc. There are several methods, including website fingerprinting attack to defeat the anonymity service of Tor technology [3,4]. In website fingerprinting attack, attackers are able to recognize the target website by comparing the traffic of the secure channel and the preprocessed traffic of the whole websites in their library. However, it is demonstrated that website fingerprinting attack based on statistical features suffers from low accuracy. On the other hand, the methods that rely on the similarity distance have faced a high time complexity that needs cluster computing to handle their demanding computation. This paper introduces a highly accurate, passive website fingerprinting attack with a low time complexity. This method benefits from the TCP/IP traffic exchanged between users and the first relay without any traffic screening, i.e. rounded packet size, without access to Tor's Cell sequence, in order to remove SENDME cells, and etc.

The paper is organized as follows: Section 2 reviews website fingerprinting attacks; Section 3 explains how the proposed method in detail; In Section 4, we will verify the output results of the proposed method and compare its performance with the existing methods in terms of time complexity and accuracy; In Section 5, analysis of the proposed method is described in detail and its major features are presented; Finally, Section 6 concludes and proposes the future works.

2. Background

Onion routing is designed to ensure anonymous and secure communications through a public network. In the onion network, like the onion's layers, messages are encapsulated in layers of encryption. The encrypted data is sent to the network relays (called onion routers) and each network relay extracts the next destination of the data by peeling its own layer. Tor is an overly low latency anonymous network based on the onion routing [2]. A Tor user installs onion proxy software (OP) and connects to the Tor directory server to get relay's policy information such as security parameter, bandwidth, IP address, etc. The OP configures three relays (Entry guard relay, Middle relay and Exit relay) in accordance with the policy information.

* Corresponding author.
E-mail addresses: h.jahani@modares.ac.ir (H. Jahani), sjalili@modares.ac.ir (S. Jalili).

http://dx.doi.org/10.1016/j.comcom.2016.05.019
0140-3664/© 2016 Elsevier B.V. All rights reserved.

The OP establishes a circuit that includes the aforementioned three relays to establish anonymous communication with a web server for a certain period of time. To this end, the OP exchanges the session key with the Entry Guard Relay (i.e. the first relay in this circuit) and makes a secure channel between them. Then, to extend the circuit further, the OP shares the session key with the second relay using the secure channel that is already made, and this process goes on until the end. Note that, in this circuit, each relay only knows the identity of its previous and next nodes (OP or relays). Since user identity is only disclosed to the Entry relay and the web server identity is revealed solely to the Exit relay, no single relay is capable to know the user and web server identities concurrently. OPs build new circuits periodically, by expiring the old used circuits and switching to a new circuit every few minutes.

A cell, with the length of 512 bytes, is the basic unit of connection in Tor. All users' messages are broken by OP into fixed size cells and sent onto the circuit. Tor uses several control cells, such as the SENDME cell to support interactions between OP and relays. These cells, similar to ACK packets at the IP-level, have an adverse effect on the results of the website fingerprinting attack. Website fingerprinting attack means that a local attacker can monitor and analyze the passing traffic flow between an OP and the first relay by accessing and counterfeiting an Entry Guard relay or penetrating it.

3. Related works in the website fingerprinting attack

The website fingerprinting attack is capable of recognizing a user target website by extracting information from the traffic patterns such as packet length, the direction of packet transmission, packet order, etc [4–6]. This attack is implemented in two phases. In the first phase, the attacker applies the same anonymity network used by users and collects instances from all the connected websites and tries to learn their patterns. In the second phase, the attacker tries to recognize the target website by monitoring the user daily traffic and matching them to the patterns resulted from the first phase. To perform this attack, two scenarios are considered: closed-world and open-world scenarios. In the closed-world scenario, it is assumed that traffic patterns of all websites are known and the attacker is going to recognize a target website from these websites. On the other hand, in the open-world scenario, just a few censored websites are known and the attacker discriminates them from a set of countless unknown websites and then tries to find out the user's visited website.

Two approaches have been introduced to prepare the features required in this attack: one approach uses statistical features of the instances and another approach uses the similarity distance of instances.

Herrmann, et al. [7] benefited from the traffic patterns of different websites in two anonymity networks of JAP and Tor. They took advantage of a machine learning technique that is based on MNB[1] and achieved an accuracy of target recognition of 20% and 3% in JAP and Tor, respectively. Shi, et al. proposed a website fingerprinting attack based on cosine similarity in 2009 [8]. They could achieve a accuracy of target recognition 50% for 20 websites. Panchenko, et al. used the traffic patterns of 775 most popular URLs and removed the ACK packets (i.e., packets with the length 52 bytes) from the traffic of the website and introduced seven features [9]. They employed the SVM[2] learning method with RBF kernel and achieved an accuracy of target recognition of 80% and 54% in JAP and Tor anonymity network, respectively.

After introduction of website fingerprinting attack, several defense methods against traffic analysis attack are proposed, such as HTTPOS and randomized pipelining [10]. Cai, et al. suggested an effective attack, which could defeat the defenses like HTTPOS and randomized pipelining [11]. They used the optimal string alignment distance (OSAD[3]) to calculate the similarity distance of each instance from others. They used SVM on their defined kernel $\mathrm{K}(\mathrm{t},\mathrm{t}') = \exp^{-\gamma \mathrm{D}(\mathrm{t},\mathrm{t}')^2}$ to reach an accuracy of target recognition of 83% in 100 websites and 74% in 775 websites for Tor anonymity network.

In 2013, Wang and Goldberg [12] used Tor cells sequence instead of the IP packets sequence and removed the SENDME cells (i.e. those control cells without important information) from Tor's traffic. In their attack, they implemented the DLD[4] by introducing new metrics based on OSAD and making an additional change to measure the similarity of instances [13]. They achieved an accuracy of target recognition of 91% for 100 websites in SVM (with the kernel used in Cai's method), meanwhile their method suffers from a high time complexity in similarity distance calculation. To resolve this problem, they applied a new distance algorithm called Fast Levenshtein-like Distance to reach a running speed which is 2000 times faster than before, despite the fact the accuracy of target recognition has dropped to 71% for 100 websites. Wang, et al performed a more extensive investigation and could increase true positive rate for 100 monitoring websites for 2% [14]. All approaches that were introduced above, operate as passive attack.

He, et al. [15] carried out active website fingerprinting against Tor anonymity and showed that Tor has not necessary resistance against this type of attack. By applying 10 requests delay (at most 1-s), they achieved a detection rate of 65% for 100 websites.

In 2015, Gu, et al. [16] proposed an active website fingerprinting based on Mahalanobis distance and classified the first page with 76% accuracy and the second page is 40.5%. They have shown that active website fingerprinting attack is a serious threat to users' privacy.

4. The proposed website fingerprinting Attack

4.1. Introduction

The proposed method is similar to other methods. It has two separate stages: a preprocessing stage and a website prediction stage. The preprocessing stage is the main core of the proposed method and consists of two activities: extracting similarity and learning classifiers. In extracting similarity, we use DFT[5] and to learn classifiers, then incorporate the SVM learning method. The most important problem at website prediction stage is speed, which can be improved by reducing the features space.

For a website fingerprinting attack, a prerequisite is collecting the required dataset. This paper uses the collected datasets in [11,12]. Each instance of the datasets which is presented by t_i in relation (1) is equivalent of a vector of transmitting packets in a secure anonymous channel to load a website,

$$t_i = d_{i1},\ \ d_{i2}, \dots d_{il}\ \ where\ d_{ik} = \pm s_{ik} \tag{1}$$

in which l is the number of packets in i^{th} instance and it is considered as the actual length of i^{th} instance.; d_k represents the k^{th} packet; s_k reflects the size of k^{th} packet and − or + sign shows the direction of the packet: the "−" sign states that the packet moves from the web server side to the user side, called incoming packet, and the "+ " sign states that the packet moves from the user side to the web server side, called outgoing packets. In other words, the positive and negative signs are used to label the direction of packet transmission.

[1] Multinomial Naïve Bayes
[2] Support vector machine
[3] OSAD is a restricted version of the DLD
[4] Damerau Levenshtein Distance
[5] Discrete Fourier Transform

Please cite this article as: H. Jahani, S. Jalili, A novel passive website fingerprinting attack on tor using fast fourier transform, Computer Communications (2016), http://dx.doi.org/10.1016/j.comcom.2016.05.019