POSTER: Evaluating Security Metrics for Website Fingerprinting

Nate Mathews
Center for Cybersecurity
Rochester Institute of Technology
nate.mathews@mail.rit.com

Mohammad Saidur Rahman
Center for Cybersecurity
Rochester Institute of Technology
saidur.rahman@mail.rit.edu

Matthew Wright
Center for Cybersecurity
Rochester Institute of Technology
matthew.wright@rit.com

ABSTRACT

The website fingerprinting attack allows a low-resource attacker to compromise the privacy guarantees provided by privacy enhancing tools such as Tor. In response, researchers have proposed defenses aimed at confusing the classification tools used by attackers. As new, more powerful attacks are frequently developed, raw attack accuracy has proven inadequate as the sole metric used to evaluate these defenses. In response, two security metrics have been proposed that allow for evaluating defenses based on hand-crafted features often used in attacks. Recent state-of-the-art attacks, however, use deep learning models capable of automatically learning abstract feature representations, and thus the proposed metrics fall short once again. In this study we examine two security metrics and (1) show how these methods can be extended to evaluate deep learning-based website fingerprinting attacks, and (2) compare the security metrics and identify their shortcomings.

KEYWORDS

Privacy, Tor, Website Fingerprinting, Information Leakage

ACM Reference Format:
Nate Mathews, Mohammad Saidur Rahman, and Matthew Wright. 2019. POSTER: Evaluating Security Metrics for Website Fingerprinting. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3319535.3363272

1 INTRODUCTION

With over 8 million users [11], the Tor anonymity system [6] is frequently chosen by privacy-conscious users to improve their privacy while minimizing usability sacrifices. Tor allows these users to conceal their location and browsing behaviors from both online tracking by web servers and local eavesdroppers. Tor is not impervious to traffic analysis attacks, however. One such attack is Website Fingerprinting (WF) [1, 8, 12–17], which allows a passive local eavesdropper to deduce information about Tor-protected traffic using traffic metadata. In a WF attack, the adversary's goal is to determine what website a Tor user has visited in a browsing session. An eavesdropper positions themselves somewhere on the link between the client and guard to perform the attack (see Figure 1).The attacker can train a machine learning classifier to distinguish between the traffic patterns of different sites of interest,

CCS '19, November 11–15, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6747-9/19/11.
https://doi.org/10.1145/3319535.3363272

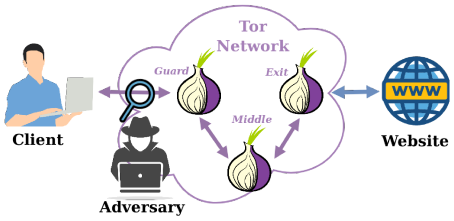

Figure 1: WF threat model.

and the most potent WF attacks use Deep Learning (DL) to achieve 98%+ accuracy [2, 15].

Several defenses have been proposed [3, 4, 7, 9, 18] in recent years to mitigate the threat of WF. These defenses change the traffic patterns produced when accessing sites by adding fake packets (*padding*) or delaying real packets. It can be difficult to accurately assess the true effectiveness of defenses that do not come with security guarantees. Evaluations are limited to testing against whatever attacks happen to be state-of-the-art at the time the defense is developed. In response, two security metrics were recently introduced—Bayes error estimate [5] and WeDFE [10]—that claim to provide a more objective measure of a defense than just attack accuracy.

In this study, we examine these metrics in more detail. Most importantly, we show how these two techniques can be extended to analyze DL-based WF attacks. Additionally, our results reveal some limitations of these metrics, indicating that further work is needed to determine the true effectiveness of WF defenses.

2 SECURITY METRICS FOR WF

Until recently, researchers evaluated the efficacy of their defense proposals by testing their simulated traffic against the state-of-the-art attacks. Consequently, defenses need to be re-evaluated whenever new attack techniques or features became available. This also limits defense evaluation to an all-or-nothing approach, where near hits and misses are discounted. Two recent works have proposed techniques to address this issue:

Bayes Error Estimation. Proposed by Cherubin [5], this metric is based on an estimation of an attacker's *Bayes error rate*. The Bayes error rate represents the lowest possible prediction error of a classifier for the target data. Simply put, if two traffic traces from two different websites have all the same features as used by any classifier to distinguish between sites, then even a perfect classifier can only guess between them. This technique expresses security in terms of $\hat{R}^*$—the Bayes error lower-bound estimate—and (ϵ,ϕ)-privacy—how close a defense is to ideal (ϵ) for a given feature representation (ϕ).

Information Leakage. Li et al. [10] proposed a technique called WeFDE to estimate the amount of information leaked (in Shannon

Table 1: Metric results for ML and DL feature experiments.

		Info. Leakage		Bayes Error		Top-1 Accuracy		Top-2 Accuracy	
		Bits	% of Max	$1-\hat{R}^*$	(ϵ,ϕ)-privacy	RF	DF	RF	DF
ML	Undefended	6.49	98.9%	90.9%	0.09	96.3%	-	97.9%	-
	WTF-PAD	6.54	99.6%	47.8%	0.52	62.5%	-	75.3%	-
	Walkie-Talkie	6.37	97.1%	45.9%	0.54	9.03%	-	89.5%	-
	Tamaraw	3.20	48.8%	28.5%	0.73	12.5%	-	21.4%	-
DL	Undefended	6.54	99.6%	97.9%	0.02	96.2%	97.1%	97.6%	98.2%
	WTF-PAD	6.48	97.8%	83.4%	0.17	81.2%	85.9%	88.5%	91.9%
	Walkie-Talkie	6.42	98.9%	72.7%	0.27	31.6%	43.8%	78.7%	98.1%
	Tamaraw	3.57	54.4%	20.3%	0.80	6.5%	7.6%	12.0%	13.2%

bits) by a defense. The WeFDE technique estimates information leakage by finding the mutual information between the distribution of sites and the information contained in the fingerprints of those sites. An advantage of WeFDE is that features can be analyzed individually.

2.1 Extending to DL

These security metrics are designed to analyze handcrafted features developed for early ML-based WF attacks. The domain, however, has recently moved to more powerful DL-based attacks that directly utilize raw traffic information. To evaluate DL attacks using these security metrics, we need to make some adjustments.

In this study we specifically examine the Deep Fingerprinting (DF) attack. The DF attack utilizes a convolutional neural network model (CNN) that can automatically learn robust feature representations from raw data. This ability is often accredited to the convolutional layers used in the early layers of the model. The outputs of convolutional layers can be thought of as the DL model's internal feature representation.

To apply the existing WF metrics to this CNN model, the learned feature representations must first be extracted. We do this by training the CNN model on a training dataset so that the convolutional filters have been learned. We then remove the classification and fully-connected layers from the model such that the the trained model returns the outputs of the last convolutional layer (see Figure 2).

3 EVALUATION

For the following experiments, we use the large datasets collected by Sirinam et al. [15]. In particular, we use their dataset containing 95 sites with 1,000 instances each for both undefended Tor and for Tor with simulated WTF-PAD [9] and Tamaraw [4] defenses. For our Walkie-Talkie (W-T) [18] evaluations, we use Sirinam's W-T dataset, which includes 900 instances each.

We run two sets of experiments between which we vary the feature representation for our data (ML or DL features). In our first set of experiments, we process data into hand-crafted features (representing ϕ) using a feature set derived from the features of CUMUL [12] and k-FP [8]. In our second set of experiments, we instead use the DL representation of the data provided by the DF attack model [15].

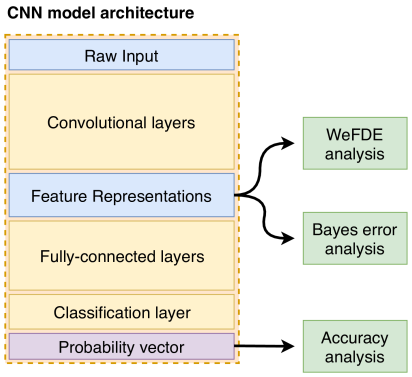

Figure 2: Process for performing metrics analysis on CNN-based DL models.

We use these experiments to compare the results of accuracy-based evaluations with that of the WeFDE and Bayes error techniques. For the accuracy evaluations, we examine feature performance for both the DF model and a Random Forest (RF) classifier.

3.1 Results

The results from our experiments are summarized in Table 1. The DL features achieve approximately a 20% improvement over ML features when used with the same RF classifier for the WTF-PAD traffic and a 23% improvement for Walkie-Talkie. As expected, the DF attack outperforms the RF classifier in nearly all settings, except for Tamaraw. This is likely due to the 5000-packet cutoff for trace length that we used for all DL experiments, removing the useful total trace length feature, since Tamaraw's high rate of dummy packets leads to very long traces.

WeFDE. When we compare the individual feature leakages to the total feature leakage, as illustrated in Figure 3, we find a surprising mismatch of results. While overall information leakage for undefended, WTF-PAD, and W-T reach near the maximum possible leakage, the individual leakage measurements show noticeably different leakage patterns. When examining the individual leakages, we see that the undefended dataset leaks on average 1.75±0.50 bits per feature value with a maximum leakage of 2.80 bits. On the other hand, the average leakage for the W-T and WTF-PAD datasets are

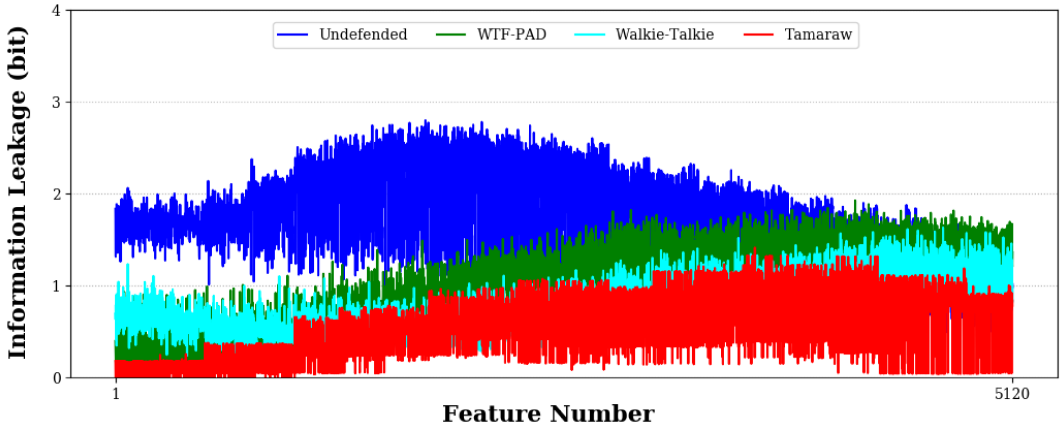

Figure 3: Individual leakage values for DF feature outputs.

0.80±0.29 and 1.01±0.45 bits, respectively, with lower maximum leakages of 1.63 and 1.92 bits. This result, particularly in the case of W-T, is much more in-line with our expectations given W-T's maximum attacker accuracy of 50%. This would seem to indicate that WeFDE's overall leakage estimates are too high and show little of the differences between defenses.

Bayes Error. The Bayes error results better correlate with the top-1 and top-2 accuracies. We notice, however, that the error bounds estimates often under-represent the threat, such as in the case of the ML feature representation for the WTF-PAD and Undefended datasets, which are both significantly lower than the actual Top-1 accuracies. It is interesting to see that Bayes error also identifies the DL-feature representation of W-T as a potentially weak defense, with an ϵ of 0.27 and a maximum attacker accuracy similar to that of the RF classifier's top-2 accuracy.

4 CONCLUSION

In this study, we examined two security metrics for evaluating WF defenses and demonstrated a method for extracting feature representations from DL-based WF attacks for use in further analysis. We found that the features learned by DL models are often more robust than their handcrafted alternatives and that these DL features produce better metric estimates when used to estimate Bayes error lower-bounds. However, we noticed that the WeFDE technique tends towards overestimation of information leakage, and the Bayes error estimation technique occasionally under represents attacker accuracy when ML features are used. These limitations demonstrate a need for additional metrics to more comprehensively evaluate WF defense strategies.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Awards No. 1722743, 1816851, and 1433736.

REFERENCES

[1] K. Abe and S. Goto. 2016. Fingerprinting attack on Tor anonymity using deep learning. In *in the Asia Pacific Advanced Network (APAN)*.

[2] Sanjit Bhat, David Lu, Albert Kwon, and Srinivas Devadas. 2019. Var-CNN: A Data-Efficient Website Fingerprinting Attack Based on Deep Learning. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (2019), 292–310.

[3] Xiang Cai, Rishab Nithyanand, and Rob Johnson. 2014. CS-BuFLO: A congestion sensitive website fingerprinting defense. In *Workshop on Privacy in the Electronic Society (WPES)*. ACM, 121–130.

[4] Xiang Cai, Rishab Nithyanand, Tao Wang, Rob Johnson, and Ian Goldberg. 2014. A systematic approach to developing and evaluating website fingerprinting defenses. In *ACM Conference on Computer and Communications Security (CCS)*. ACM, 227–238.

[5] Giovanni Cherubin. 2017. Bayes, not Naïve: Security bounds on website fingerprinting defenses. *Privacy Enhancing Technologies Symposium (PETS)* (2017).

[6] Roger Dingledine, Nick Mathewson, and Paul F. Syverson. 2004. "Tor: The second-generation Onion router". In *USENIX Security Symposium*. USENIX Association, 303–320.

[7] Kevin P. Dyer, Scott E. Coull, Thomas Ristenpart, and Thomas Shrimpton. 2012. Peek-a-Boo, I still see you: Why efficient traffic analysis countermeasures fail. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 332–346.

[8] Jamie Hayes and George Danezis. 2016. k-fingerprinting: A robust scalable website fingerprinting technique. In *USENIX Security Symposium*. USENIX Association, 1–17.

[9] Marc Juarez, Mohsen Imani, Mike Perry, Claudia Diaz, and Matthew Wright. 2016. Toward an efficient website fingerprinting defense. In *European Symposium on Research in Computer Security (ESORICS)*. Springer, 27–46.

[10] Shuai Li, Huajun Guo, and Nicholas Hopper. 2018. Measuring information leakage in website fingerprinting attacks and defenses. In *ACM Conference on Computer and Communications Security (CCS)*.

[11] Akshaya Mani, T Wilson-Brown, Rob Jansen, Aaron Johnson, and Micah Sherr. 2018. Understanding Tor Usage with Privacy-Preserving Measurement. *arXiv preprint arxiv.org/abs/1809.08481* (2018).

[12] Andriy Panchenko, Fabian Lanze, Andreas Zinnen, Martin Henze, Jan Pennekamp, Klaus Wehrle, and Thomas Engel. 2016. Website fingerprinting at Internet scale. In *Network & Distributed System Security Symposium (NDSS)*. IEEE Computer Society, 1–15.

[13] Mohammad Saidur Rahman, Payap Sirinam, Nate Matthews, Kantha Girish Gangadhara, and Matthew Wright. 2019. Tik-Tok: The Utility of Packet Timing in Website Fingerprinting Attacks. *arXiv preprint arXiv:1902.06421* (2019).

[14] Vera Rimmer, Davy Preuveneers, Marc Juarez, Tom Van Goethem, and Wouter Joosen. 2018. Automated Website Fingerprinting through Deep Learning. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society.

[15] Payap Sirinam, Mohsen Imani, Marc Juarez, and Matthew Wright. 2018. Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning. *ACM Conference on Computer and Communications Security (CCS)* (2018).

[16] Tao Wang, Xiang Cai, Rishab Nithyanand, Rob Johnson, and Ian Goldberg. 2014. Effective attacks and provable defenses for website fingerprinting. In *USENIX Security Symposium*. USENIX Association, 143–157.

[17] Tao Wang and Ian Goldberg. 2013. Improved website fingerprinting on Tor. In *ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 201–212.

[18] Tao Wang and Ian Goldberg. 2017. Walkie-talkie: An efficient defense against passive website fingerprinting attacks. In *USENIX Security Symposium*. USENIX Association, 1375–1390.