Reading Response 1: Evaluating Evaluations

CSEC 720 (2225) - Mehul Sen

"k-fingerprinting: a Robust Scalable Website Fingerprinting Technique" by J. Hayes and G. Danezis and "Automated Web Fingerprinting through Deep Learning" by V. Rimmer et al. are two research papers that both focus on Website Fingerprinting (WF) through the use of deep learning algorithms and consider datasets consisting of traffic routed through TOR.

Whilst covering similar topics, they both differ in the way they go about evaluating their WF methods. An important difference between the evaluations for both these papers was the varying focus they had. The evaluation performed in the k-fingerprinting paper focuses on comparing the newly proposed WF technique with other previously known WF techniques. It puts a greater emphasis on evaluating the technique's performance in an open-world scenario consisting of a more realistic ratio of unmonitored to monitored sites in their datasets. The Automated WF paper focuses on comparing the use of WF using three popular deep learning models with automated features with more traditional WF techniques that involved hand-picked features. Its evaluation also emphasizes setting up ideal test conditions to accurately compare the WF techniques. Therefore, the majority of their evaluation was on a closed-world dataset and their open-world dataset had a balanced proportion (50%-50% ratio) of monitored and unmonitored sites.

There are several other differences between the two papers and the datasets their evaluations are based on. While J. Hayes and G. Danezis' paper evaluates the results on datasets that consist of both normal traffic and encrypted traffic containing standard webpages and hidden services over TOR, Rimmer et al. only use datasets comprised of standard webpages encrypted over TOR. Another interesting difference between the two papers is that while J. Hayes and G. Danezis mention that time would negatively impact the WF attack, they do not evaluate the effect time has on the attacks performed. Rimmer et al. evaluate the concept drift and the change in accuracy of their WF techniques zero, three, ten, twenty-eight, forty-two, and fifty-six days after their creation.

The evaluation results shown by the k-fingerprinting paper are highlighted using tables and graphs mostly containing the number of pages trained, True Positive Rate(TPR), False Positive Rate(FPR), and Bayesian Detection Rate(BDR). While those shown by the Automated WF paper are highlighted using tables and graphs mostly containing the number of traces involved, accuracy percentage, and loss.

After going through both papers, I believe that "Automated Web Fingerprinting through Deep Learning" by Rimmer et al. has a better evaluation than the k-fingerprinting paper. This is due to the more comprehensive information and clearer comparisons between the different WF techniques made by that paper. I believe that this section could further be elaborated and improved on by including additional datasets comprising a much more realistic proportion of monitored and unmonitored sites as well as the inclusion of TOR hidden services in the open-world dataset.