Assessing the Feasibility of Machine Learning to Detect Network Covert Channels

Diogo Barradas
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
diogo.barradas@tecnico.ulisboa.pt

ABSTRACT

Several tools allow for the creation of covert channels in the Internet by concealing data in the application layer of encrypted protocols. One of the main indicators of the quality of such tools is a measure of the difficulty involved in identifying a flow carrying covert data. Thus far, such assessments made use of a small space of ad-hoc traffic characteristics and classification techniques. Recent advances on machine learning (ML) techniques open the possibility for analyzing multiple characteristics of network flows which have not yet been considered for the detection of covert traffic. In the one hand, it is not clear that effective ML techniques can be applied in a scalable way to detect covert traffic among large amounts of legitimate Internet traffic. On the other hand, the developers of such tools may be able to bias the results of learning algorithms to thwart an adversary's efforts aimed at detecting covert channels. This proposal aims to study the advantages and potential drawbacks of the use of ML techniques for detecting covert channels.

1 INTRODUCTION

Nowadays, powerful adversaries such as governments and corporations possess the technical capabilities to exercise an active monitoring and control over the Internet traffic flowing across a country's borders. To provide the ability to transfer data away from prying eyes (e.g., for censorship circumvention purposes), multiple systems have been developed with the purpose of creating covert channels which enable the transmission of concealed data over the Internet. Lately, such systems take advantage of encrypted protocols which content an adversary is unable to snoop into through deep packet inspection techniques. A state-of-the-art approach for the design of such systems, named *protocol tunneling*, focus the embedding of concealed data directly into the application layer of multiple encrypted protocols [3, 5, 6].

Albeit adversaries are prevented from directly inspecting the content of encrypted transmissions, they may still be able to detect subtle differences in the packet traces of a protocol when it is used for carrying covert data. Such differences can be uncovered through the application of strictly passive methods (e.g., by observing the length or inter-arrival delay of transmitted network packets) or by analyzing the protocol's behavior in response to active network manipulations (e.g., the loss or reorder of packets) [2, 4].

Thus, an important property that all protocol tunneling systems strive to achieve is *unobservability*. A covert channel is deemed unobservable if an adversary that is able to scan any number of streams is unable to distinguish those that carry a covert channel from those that do not [4]. In practice, systems that provide a high degree of unobservability prevent an adversary from flagging a large fraction of covert flows while flagging a low amount of regular traffic. Erroneously flagging legitimate traffic as covert flows may be unwise for an adversary in most practical settings. For example, a censor that aims at blocking covert channels may be unwilling to block large fractions of legitimate traffic, as targeted protocols may be key for the economy of the censor's regime. Also, law-enforcement agencies may be unwilling to risk to falsely flag legitimate actions of citizens as criminal activity.

2 OPEN ISSUES

We identify several open issues in the assessment of unobservability.

Ad-hoc evaluation. So far, the unobservability of protocol tunneling systems has been performed in an *ad hoc* fashion with the application of independently built supervised classifiers based on the similarity of packet-level frequency distributions (e.g., packet lengths or inter-arrival delay). Similarity functions employed in unobservability assessments include Pearson's χ^2 test (χ^2) [7], the Kullback-Leibler divergence (KL) [8], or the Earth Mover's Distance (EMD) [2]. However, in spite of providing no information about which ranges of the feature space are most affected when covert channels are embedded in a protocol, it is debatable whether this classification approach can accurately distinguish between legitimate and covert flows when compared to a vast universe of machine learning (ML) techniques existing today. Moreover, the use of multiple attributes of network flows, which have previously yielded successful results when applied to similar domains (e.g., website fingerprinting [1]), have not been considered in this context.

Labeled data requirements. Current methodologies for the evaluation of the unobservability of covert channels assume that an adversary has unlimited access to tools providing the ability to generate covert channels. This allows an adversary to synthesize a dataset and to train a classifier for distinguishing between legitimate and covert data transmissions. However, such assumption presents multiple challenges for an adversary. Firstly, the adversary is required to build tailored classifiers for different tools which may use the same protocol as a cover for concealed data transmissions [2, 7]. Secondly, it is debatable whether an adversary would have an expedite access to such tools. Lastly, even if the adversary is assumed to possess a given tool, it is expected to spend a non-negligible time in synthesizing covert data samples for building a model. Overcoming such challenges opens a timeframe where the covert traffic generated by a given system would remain undetected.

Thus far, related literature has overlooked the possibility for adversaries to build a representative model of legitimate traffic and to flag all traffic exhibiting deviating features. If state-of-the-art semi-supervised ML techniques reveal promising results on the detection of covert traffic, it is possible that adversaries will be able to avoid the burden of synthesizing covert traffic datasets. This rationale may be further extended to unsupervised ML settings, where neither legitimate nor covert data labels are necessary.

Traffic dataset representativeness. Synthesizing legitimate traffic datasets raises the concern of whether the resulting samples are representative of a larger universe of traffic expected in the wild. For building a more realistic dataset, an adversary could benefit from its privileged position in the network and collect all data originated by a given protocol. However, the very existence of protocol tunneling tools enabling the creation of covert channels over these protocols makes it hard for an adversary to know, before-hand, which data samples correspond either to legitimate or covert traffic. In practice, training data may be polluted with adversarial examples which will affect the accuracy of the resulting models.

Lack of theoretical reasoning. Multiple protocol tunneling systems embed covert data by encoding it into the application layer of a protocol, e.g. crafted images in a videoconferencing application. So far, the tuning of such encoding mechanisms is exclusively based on empirical evidence obtained through black box experimentation. Such approaches fail to provide a theoretical formulation which allows a clear reasoning over the amount of covert information that can be encoded in such channels while maintaining unobservability.

3 PROBLEM STATEMENT

With this proposal, we aim at understanding whether the unobservability claims of protocol tunneling systems hold against an adversary making use of state-of-the-art ML techniques. For meeting this goal, we first seek to understand which underlying characteristics of traffic are more advantageous for an adversary to identify covert executions of a given protocol. Then, we seek at studying the viability of ML techniques for identifying covert channels, both when an adversary is assumed to possess fully labeled data or when it is totally or partially deprived of labeled data. as an additional step for fulfilling our goal, we seek to assess the limitations imposed over the scalability of ML techniques, sprouting from the vast diversity of network flows required to analyze, and the limitations imposed over the prediction performance of ML models due to the inclusion of adversarial training data. Lastly, we aim at providing a theoretical bound for the maximum amount of covert data that can be embedded in a carrier protocol, while achieving perfect unobservability, despite the use of increasingly sophisticated ML traffic classifiers for detecting protocol tunneling systems.

4 RESEARCH PLAN

The next paragraphs detail our research plan.

1. Perform a comparison over the classification performance of similarity-based classifiers and state-of-the-art machine learning techniques when identifying covert channels. We shall assess the performance of decision tree-based algorithms in identifying covert channels, and compare it to the results obtained by current similarity-based classifiers. Additionally, we will measure the memory and computational costs for training and performing predictions with such models. Thanks to the easy interpretability of decision tree-based models, we shall analyze the importance of features in order to identify which attributes of the feature space are more affected by covert data embedding. The acquisition of such fine-grained information could provide useful both for adversaries aiming at detecting covert channels, and for the maintainers of covert channels-generating systems in order to create robust covert data encoding mechanisms. As a simplification, we shall synthesize traffic datasets in laboratory.

2. Perform the identification of covert channels while avoiding to possess fully labeled datasets. We shall assess the feasibility of detecting covert channels while we assume an adversary to be partially or fully deprived of labeled data. To this end, we shall build one-class classification models for identifying covert channels while characterizing legitimate traffic alone, and use unsupervised learning models for identifying covert traffic in the complete absence of labeled data. We shall synthesize legitimate/covert traffic in laboratory settings for conducting our experiments.

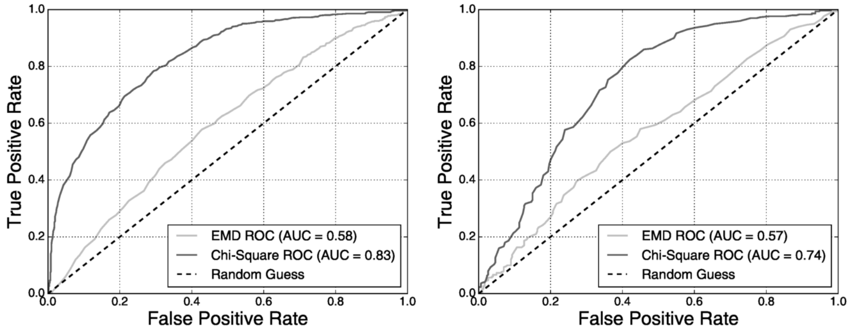
3. Analyze the influence of adversarial samples in the ability of uncovering covert channels. ML models may be retrained for reflecting data observed during a given execution period. We shall formulate the task of detecting covert channels as a problem of adversarial ML, where the maintainers of protocol tunneling tools are assumed to pollute the training data of an adversary with crafted samples aimed at biasing the predictions of an adversary's models. We shall evaluate the potential of dataset pollution as a countermeasure for the detection of covert channels, and the trade-offs involved for an adversary in order to sanitize such models.

4. Establish theoretical bounds to unobservability. The ultimate goal of our research plan is that of providing theoretical foundations for the unobservability of protocol tunneling systems. To this end, we shall combine the fine-grained insights gathered through previous ML experimentation with a thorough analysis of the characteristics of the data channels used by protocol tunneling systems. Then, we aim at building models which can formally dictate limits on the amount of covert data that can be embedded in carrier protocols while achieving perfect unobservability, even in the face of adversaries making use of sophisticated ML techniques.

5 CURRENT STATUS

We fulfilled the first stage of the research plan by conducting an experimental study over the unobservability properties of three state-of-the-art protocol tunneling systems which create covert channels in multimedia streaming protocols. Facet [7] allows clients to watch arbitrary videos by replacing the audio and video feeds of Skype calls. Facet approximates the traffic patterns of regular videocalls by re-sampling the audio frequency and overlaying the desired video in a fraction (s) of each video frame. CovertCast [8] modulates the content of web pages into matrix-like images which are distributed via live-streaming platforms such as YouTube. Matrix images are parameterized by a cell size (adjacent pixels with a given color), the number of bits encoded in each cell (represented with a color), and the colored matrix framerate. DeltaShaper [2] tunnels TCP/IP traffic by modulating it into colored matrices which are transmitted through a bi-directional Skype videocall. A colored matrix is overlayed in a fraction of the call screen. This overlay is parameterized according to the observed network conditions to increase the covert channel's unobservability, and respects the ⟨payload frame area, cell size, number of bits, framerate⟩ tuple.

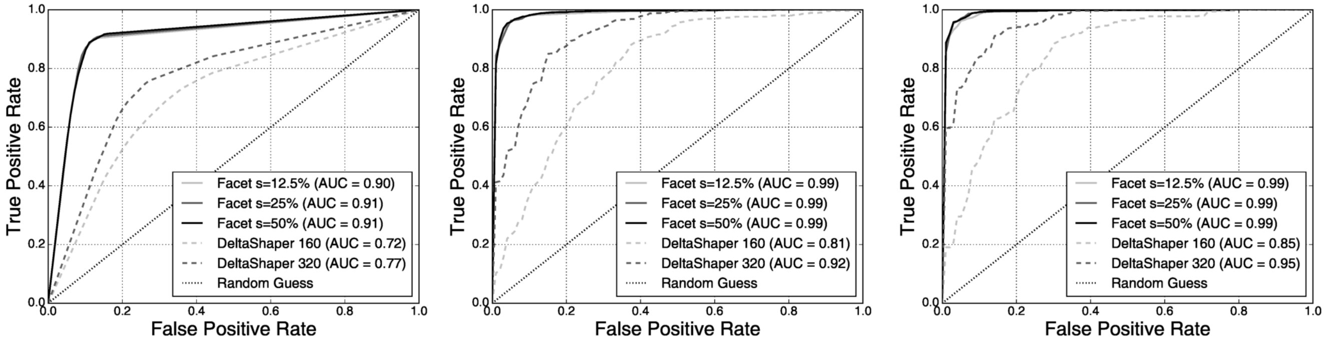
In our study, we evaluated the unobservability of each system against each of the known similarity-based classifiers ($\mathcal{X}^2$, KL, and EMD). Then, we evaluated the unobservability of the same systems by resorting to three different decision tree-based algorithms: a



(a) Facet *s*=50%

(b) DeltaShaper ⟨320 × 240, 8 × 8, 6, 1⟩.

Figure 1: ROC curve for the $\mathcal{X}^2$ and EMD classifiers when identifying Facet and DeltaShaper traffic.



(a) Decision Tree.

(b) Random Forest.

(c) XGBoost.

Figure 2: ROC curves of the decision tree-based classifiers when identifying Facet and DeltaShaper traffic.

simple C4.5 decision tree, and two prominent decision tree ensemble methods (Random Forest and XGBoost). Akin to the feature set used in similarity-based classifiers, we used the quantized frequency distribution of packet lengths as features. However, we exploited the relevance of particular ranges of the feature space by feeding this feature set to decision tree-based classifiers.

The outcomes of our experiments reveal several interesting findings. First, we find that CovertCast can be detected with few false positives (FP) when resorting to existing similarity-based classifiers. Second, we conclude that the $\mathcal{X}^2$ classifier outperforms all other classifiers of its kind in the task of detecting covert channels. As an example, Figure 1 depicts the ROC curves for the $\mathcal{X}^2$ and EMD classifiers when detecting Facet *s*=50% and DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1\rangle$ traffic. Albeit we can observe that $\mathcal{X}^2$ outperforms EMD for this kind of assessment, we also observe that the $\mathcal{X}^2$ classifier fails to detect a large amount of covert flows while sustaining a low FP rates (e.g. to block 90% of all Facet *s*=50% traffic, the $\mathcal{X}^2$ classifier erroneously tags 45% of legitimate connections as covert traffic). We do not show a ROC curve for the KL classifier as it is not adjustable by an internal threshold. However, its accuracy was found to be inferior to that of a simpler version of $\mathcal{X}^2$ with no adjustable threshold. Lastly, our results suggest that decision tree-based classifiers largely defy the previous unobservability claims of existing multimedia protocol tunneling systems. This finding is supported by the results in Figure 2, which show the performance of different decision-tree based classifiers when detecting different configurations of Facet and DeltaShaper. For instance, 90% of all Facet *s*=50% traffic can be detected with just 2% FP rate. Furthermore, we found that the accurate detection of different systems is closely tied to disparate ranges of quantized packet lengths.

In order to accomplish the second stage of our research plan, we evaluated each systems resorting to One-Class SVMs and autoencoders, two state-of-the-art ML techniques able to perform one-class classification. Our results indicate that One-Class SVMs perform poorly on identifying the covert channels produced by these systems, while autoencoders show promising results for covert traffic detection. We have also evaluated the unobservability of the above systems resorting to Isolation Forest, an unsupervised learning algorithm. The outcome of this experiment reveals that an adversary has no advantage in using Isolation Forest for the detection of covert traffic. Thus, our findings suggest that the existence of manually labeled samples is a requirement for the successful detection of multimedia protocol tunneling covert channels.

ACKNOWLEDGMENTS

This work was partially supported by national funds through Instituto Superior Técnico, Universidade de Lisboa, and Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 (INESC-ID).

REFERENCES

[1] Khaled Al-Naami, Swarup Chandra, Ahmad Mustafa, Latifur Khan, Zhiqiang Lin, Kevin Hamlen, and Bhavani Thuraisingham. 2016. Adaptive encrypted traffic fingerprinting with bi-directional dependence. In *Proc. of ACSAC*.

[2] Diogo Barradas, Nuno Santos, and Luís Rodrigues. 2017. DeltaShaper: Enabling Unobservable Censorship-resistant TCP Tunneling over Videoconferencing Streams. In *Proc. of PETS*.

[3] Chad Brubaker, Amir Houmansadr, and Vitaly Shmatikov. 2014. CloudTransport: Using Cloud Storage for Censorship-Resistant Networking. In *Privacy Enhancing Technologies*.

[4] John Geddes, Max Schuchard, and Nicholas Hopper. 2013. Cover Your ACKs: Pitfalls of Covert Channel Censorship Circumvention. In *Proc. of ACM CCS*.

[5] Bridger Hahn, Rishab Nithyanand, Phillipa Gill, and Rob Johnson. 2016. Games without frontiers: Investigating video games as a covert channel. In *Proc. of IEEE EuroS&P*.

[6] Amir Houmansadr, Thomas J. Riedl, Nikita Borisov, and Andrew C. Singer. 2013. I want my voice to be heard: IP over Voice-over-IP for unobservable censorship circumvention.. In *Proc. of NDSS*.

[7] Shuai Li, Mike Schliep, and Nick Hopper. 2014. Facet: Streaming over Videoconferencing for Censorship Circumvention. In *Proc. of ACM WPES*.

[8] Richard McPherson, Amir Houmansadr, and Vitaly Shmatikov. 2016. CovertCast: Using Live Streaming to Evade Internet Censorship. In *Proc. of PETS*.