Poster: Zero-day Botnet Domain Generation Algorithm (DGA) Detection using Hidden Markov Models (HMMs)

Yu Fu, Lu Yu, Richard Brooks *Senior Member, IEEE*
Holcombe Department of Electrical and Computer Engineering
Clemson University, Clemson, SC, 29634, USA
fu2, lyu, rrb@g.clemson.edu

I. Introduction

Domain Generation Algorithms (DGAs) are widely used by modern botnets to keep botnets hidden from security personnel. They can generate a large number of domains, which are used as rendezvous between botmasters and bots. The advantage of DGA use is not only its ability to conceal the botnet command and control (C&C) node, but also the security vendors' inability to destroy botnet since they need to register all possible DGA domain names beforehand to sinkhole botnet communication.

Yadav et al. [1] proposed three detection methods, Kullback-Leibler (KL) distance, Edit distance (ED), and Jaccard Index (JI), which achieved up to 100% detection rate and 2.5% false-positive rate. Our previous work [2] developed DGAs that use Hidden Markov Models (HMMs) and Probabilistic Context-Free Grammars (PCFGs) that effectively evade these three detection methods. In contrast to that work, this work uses the same HMMs to detect botnet DGAs.

In this work, we proposed the zero-day DGA detection method using Hidden Markov Models (HMMs). The idea is to compare a DGA-generated domain name with the HMM that represents the lexical features of the legitimate domain names. By calculating the probability that the given domain name is generated by the HMM, we can decide how likely it is a DGA-generated domain name. A Receiver Operating Characteristic (ROC) curve is used to find the optimal threshold for the decision process.

Unlike other DGA detection techniques, our method doesn't require prior knowledge of botnet DGAs. Therefore, it is suitable for detecting previously unknown botnet DGAs. We use game theory analysis to model the combat between botmasters and security personnel, which will help security personnel develop effective botnet detection strategies.

II. Methods & Results

Fig. 1 shows the procedure of the zero-day HMM-based DGA detection method. Given a test domain name, we calculate the probability that the test domain name is generated by the legitimated HMM as P. We determine whether it should be labeled as malicious by comparing P with a threshold. Then we draw ROC curves to visualize the pairs of True Positive Rate (TPR) and False Positive Rate (FPR) to find the optimal threshold as the best operating point for future botnet detection. Since our HMM-based DGA detection does not require training datasets, it is suitable for detecting previously unknown botnet DGAs.

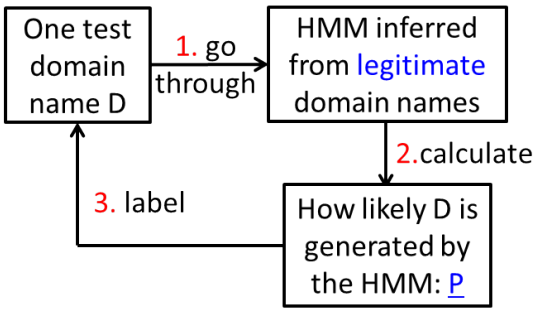

Fig. 1: HMM detection method for new DGAs

We tested all DGA detection methods on 5 botnet DGAs (pushdo, tinba, ramdo, srizbi and rovnix). Fig. 2 shows the ROC curves of the detection performance of HMM, KL, ED and JI detection methods. HMM detection gives a good detection rate with at least 74.2% TPR and at most 28.1% FPR on all DGAs. It shows that HMM detection method works better than KL, ED or JI when detecting new DGAs.

Also, we modelled the game between the botmaster (the one who uses DGAs to evade detection) and security personnel (the one who detects DGAs) as a Two-Person Zero-Sum (TPZS) game [3], because they have conflicting interests. If we can find the DGAs (or detection methods) that the botmaster (or the security personnel) tends to choose for the optimal interest, we can have a better understanding of the arm race between the two parties [4]. Table I shows the distance from the best operating point to perfect detection point (0,1) of all DGAs under 4 detection methods. Game theory analysis results show that, to optimize DGA detection for the current generation of DGAs, security personnel should use the HMM detection method, and botmasters should choose the pushdo DGA.

III. Conclusion & Future Work

We proposed an HMM-based detection method and compared it with 3 other detection methods(KL/ED/JI) on 5 botnet DGA datasets. The experiment results show that the HMM

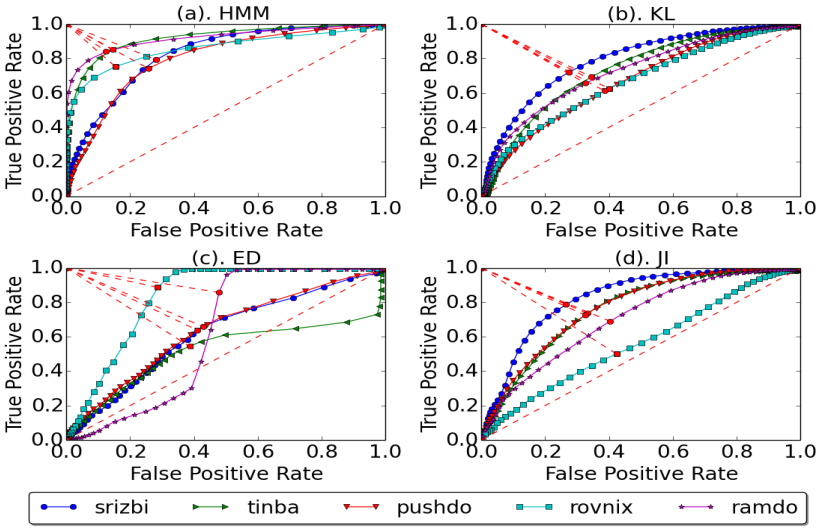

Fig. 2: ROC curves of (a) HMM, (b) KL, (c) ED, (d) JI detection on 5 DGAs

TABLE I: Game theory analysis

Player II: Bomaster (maximizer) \ Player I: security personnel (minimizer)	HMM	KL	ED	JI	min
srizbi	0.12	0.15	0.30	0.12	0.12
tinba	0.04	0.21	0.36	0.18	0.04
pushdo	0.13	0.30	0.30	0.18	0.13
rovnix	0.08	0.30	0.10	0.43	0.08
ramdo	0.04	0.22	0.25	0.26	0.04
max	0.13	0.30	0.36	0.43	

method outperforms the KL/ED/JI detection in detecting previously unknown DGAs.

To our best knowledge, this is the first time that an HMM is used for DGA detection. The results are promising. It is helpful for almost any DGA detection system using lexical features. We hope this paper can help security personnel improve the current detection techniques, and take down botnets in a larger scale.

Future work includes: (1) test HMM detection on larger datasets; (2) try different parameters to improve the HMM detection methods; and (3) test HMM detection against other string metrics.

REFERENCES

[1] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan, "Detecting algorithmically generated domain-flux attacks with dns traffic analysis," *IEEE/Acm Transactions on Networking*, vol. 20, no. 5, pp. 1663–1677, 2012.

[2] Y. Fu, L. Yu, O. Hambolu, I. Ozcelik, B. Husain, J. Sun, K. Sapra, D. Du, C. T. Beasley, and R. Brooks, "Stealthy domain generation algorithms (dgas)," *IEEE Transactions on Information Forensics & Security*, 2017.

[3] J. Von Neumann and O. Morgenstern, "Theory of games and economic behavior," *Bull. Amer. Math. Soc*, vol. 51, no. 7, pp. 498–504, 1945.

[4] Y. Fu, B. Husain, and R. R. Brooks, "Analysis of botnet counter-countermeasures," in *Proceedings of the 10th Annual Cyber and Information Security Research Conference*. ACM, 2015, p. 9.