Poster: Botmaster Detection using Passive DNS and Network Flow

Bo Hu, Takaaki Koyama, and Jun Miyoshi
Secure Platform Laboratories, NTT
Tokyo, Japan
{hu.bo, koyama.takaaki, miyoshi.jun}@lab.ntt.co.jp

Kenji Takahashi
Global Research, NTT Security
Palo Alto, USA
kenji.takahashi@nttsecurity.com

***Abstract*—For tracking down the root cause of botnets, we propose a botmaster detection method using passive DNS and network flows. The proposed method extracts C&C server groups with DNS records and detect botmaster of each group by applying unsupervised learning to network flows.**

Keywords—botnet; botmaster; network flow; passive DNS; machine learning

I. INTRODUCTION

In recent years, malware and botnets become a serious cyber security problem from both of economic and social perspectives. Industry and academic institutions have many remarkable achievements to detect the command and control (C&C) channel which plays the key function for attackers to remotely control infected hosts. Especially, machine learning base network flow analysis is considered as one of effective approaches to detect characteristic behaviors in C&C channels without inspecting packet payloads for privacy considerations [1][2][3].

However, as shown in Fig.1, botnet related techniques are also growing and evolving at the same time. Fast-flux, Domain Generation Algorithm (DGA) and hierarchical botnet structure, those techniques help attackers deploy multi C&C servers and bind IP addresses with massive randomly generated domain names so that attackers can change bot entry points flexibly as well as they can keep the control of bots even though some of domains and IP addresses were blocked [4][5].

To address these counter measures of the attacker side, we take a novel approach to eliminate the root cause by tracing and tracking down botmasters. This approach has led us to invent a technique to detect botmasters from limited sets of network flows and DNS data in the Internet. The major contributions of this approach are to overcome the following limitations:

- Lack of complete flows: It is impractical to obtain complete flows between any two hosts in the Internet due to the device performance limitation. Alternatively, packets are sampled in one out of several thousand at routers and switches, namely, most of packets between a botmaster and a C&C server cannot be observed in flows.
- No pre-knowledge: There are no labels of botmasters for supervised machine learning which can achieve high detection accuracy by training classifiers with labels.

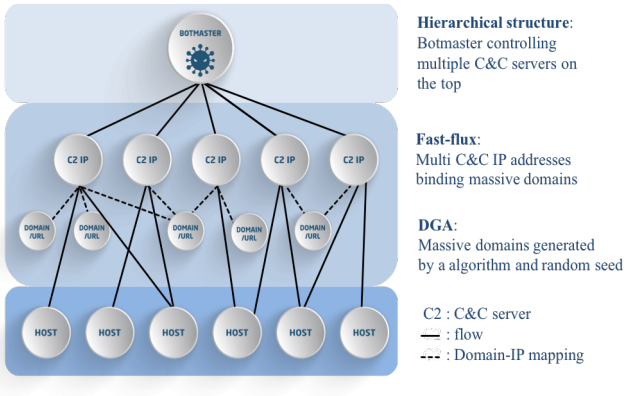

Fig. 1. Example of botnet infrastructure

In this research, we focus on hierarchical and centralized botnet infrastructures and propose a botmaster detection method. This proposal groups multiple C&C servers with passive DNS records, and detect the botmaster by applying unsupervised learning to network flows. We analyze the interactions between a host and a whole C&C server group rather than a single C&C server to obtain more flow data and new features against the lack of complete flows, and perform unsupervised learning to detect the outlier host behaving differently from others as the botmaster without any labels of botmasters.

In some case, a botmaster may deploy upper C&C servers to indirectly control lower C&C servers in a more hierarchical way. For the convenience and brevity, this paper excludes that case. However, it can be expected to detect these upper servers in the same fashion hierarchically with our proposal.

In this paper, Section II describes related works to detect botmaster, Section III describes details of our proposal using passive DNS and network flow, Section IV shows primitive results, and Section V concludes results and briefs future works.

II. RELATED WORK

The botmaster detection is a challenging area, and several previous researches were proposed [6][7][8]. Nunnery et al. reported a hierarchical spamming botnet structure and analyzing its botmaster by using the C&C server images obtained from affected hosting providers [6]. This paper provides deep insights to understand the structure and protocol of a structured botnet infrastructure. However, obtaining server images is not too often in the real world. Ramsbrock et al. proposed a packet

watermarking method to detect traffic between the C&C server and botmaster even though traffic is encrypted or passes through several stepping stones [7]. However, this method requires a huge amount of monitor points deployed across the Internet to mirror traffic without any sampling, which are difficult to meet real-world network requirements. Mizoguchi et al. proposed a communication pattern sharing framework to trace botmaster [8]. However, this framework has not considered sampled flows, and requires operators to clarify all the IP address connected to a C&C server for judging an unknown node as the botmaster.

III. Proposal

We focus on two types of data: One is the passive DNS record which shows DNS resolution history, and the other is the (sampled) network flow which shows statistics of each flow aggregated by protocol, source/destination IP addresses and source/ destination port numbers during a desired period.

As shown in Figure 2, the proposed method A) utilizes passive DNS records to group C&C servers which share the same domains, B) extract all the hosts communicated to each C&C group from network flows, and finally, C) generates flow features for each host and applies unsupervised learning to detect the botmaster which behave in an extremely different way from infected bots. Details for each step are shown as the followings.

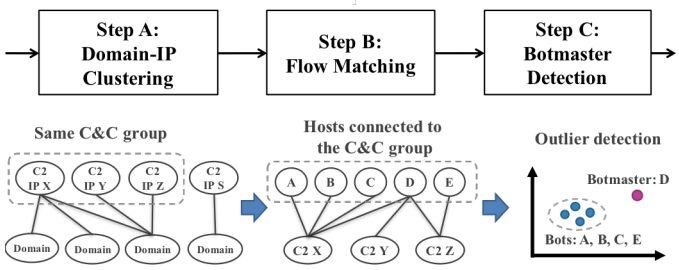

Fig. 2. Overview of proposal

In step A, we use third-party C&C domain/IP blacklists and Passive DNS records. After extracting all the DNS records including malicious entities, we build a graph that consists of nodes representing domains and IP addresses and edges representing DNS resolution history, and extract disjoint unions on the graph as different C&C groups.

In step B, after obtaining C&C IP address groups, the proposed method applies flow matching to extract the hosts communicated to each group in bi-direction flows. By analyzing communication patterns between each candidate host and the whole C&C server group, more flows and features can be expected to achieve higher accuracy through sampled flows.

In step C, we assume that the botmaster manages the availability of C&C servers, harvests data from and send commands to C&C servers. So, the botmaster may have higher availability, and connect to more C&C servers both in total and a short period than a bot. Based on the assumptions, the proposal generates the following flow features for each host: 1) unique C&C IP addresses overall, 2) unique C&C IP addresses per 5 minutes, and 3) number of timeslot observed. After that, the proposed method applied unsupervised learning to detect the outlier host. Here, we apply DBSCAN algorithm [9].

IV. Primitive Experiment

We performed step A with blacklists from third party venders and passive DNS records from VirusTotal [10]. With 299 C&C IP addresses and 31737 DNS records, 130 C&C groups were generated, and 20 groups include more than 3 C&C servers. Details of top 3 large groups are shown in Table 1. Also, we have conducted steps B and C with a limited set of sampled network flow data. Even with the sampled data, we detected tens of thousands of bots that talked to the largest group and identified one botmaster candidate. To confirm the results, we need to use larger data sets and closely investigate the candidates with human experts.

TABLE I. Domain and IP of Each C&C Server Group (Top 3)

Group	# of C&C domain	# of C&C IP
No.1	7980	84
No.2	2800	10
No.3	63	5

V. Conclusion and Future Work

In this paper, we proposed a method to detect botmaster with passive DNS records and network flows. Since the input data are reasonable to be collected from the real world, the proposal is considered as a highly reliable method to track down the botmaster. In primitive experiments, we obtained 20 C&C server groups, and detect 1 botmaster candidate of the largest group. In future, we will apply more flows to extract botmaster candidates. Moreover, since there is no pre-knowledge of botmaster, more systematic validation method also needs to be discussed.

References

[1] B Li, J Springer, G Bebis and MH Gunes, “A survey of network flow applications,” Journal of Network and Computer Applications, Volume 36, Issue 2, March 2013, Pages 567–581

[2] G Gu, R Perdisci, J Zhang and W Lee, “BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection,” USENIX Security Symposium, San Jose, USA., July 2008.

[3] L Bilge, D Balzarotti, W Robertson and E Kirda, “Disclosure: detecting botnet command and control servers through large-scale NetFlow analysis,” ACSAC '12, Pages 129-138, Orlando, USA, December 2012.

[4] T Holz, C Gorecki, K Rieck and FC Freiling, “Measuring and Detecting Fast-Flux Service Networks,” NDSS Symposium, San Diego, USA, September 2008.

[5] M Antonakakis, R Perdisci and Y Nadji, “From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware”, USENIX security, Bellevue, USA, August 2012.

[6] C Nunnery, G Sinclair and BBH Kang, “Tumbling Down the Rabbit Hole: Exploring the Idiosyncrasies of Botmaster Systems in a Multi-Tier Botnet Infrastructure,” LEET, San Jose, USA, April 2010.

[7] D Ramsbrock, X Wang and X Jiang, “A First Step towards Live Botmaster Traceback,” RAID 2008, Cambridge, USA, September 2008.

[8] S Mizoguchi, K Takemori and Y Miyake, “Traceback Framework against Botmaster by Sharing Network Communication Pattern Information,”, IEEE IMIS 2011, Seoul, Korea, June 2011

[9] M Ester, HP Kriegel, J Sander, X Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” KDD, Portland, USA, August 1996.

[10] VirusTotal, www[.]virustotal[.]com