Inspecting DNS Flow Traffic for Purposes of Botnet Detection

Vojtěch Krmíček, GEANT3 JRA2 T4 Internal Deliverable

2011

Abstract

The goal of this report is to examine DNS IP flow traffic and its relation to the botnet presence in the observed network. We focus on the DNS traffic generated by the botnets in general and inspect existing botnet detection methods using DNS data. DNS traffic in the form of IP flow data is studied and the possibility to use DNS detection methods on the IP flow data is discussed. An analysis of DNS traffic from backbone network is presented and directions for future work are provided.

1 Introduction

The botnets represent one of the biggest Internet threats nowadays and they are still more frequent and spread all over the world. They are developed rapidly due to the huge hidden economics behind them [11], therefore their detection and defense against them is more difficult. The botnet research is and will be the important task for nowadays Internet security researchers.

Malicious traffic is traditionally detected by deep packet inspection: the payload is searched for signatures of known attacks. However, this is very resource-intensive task and scalability is a growing problem in current multigigabit networks. Contrary, an intrusion detection based on an analysis of network flows (usually in the *NetFlow* format) scales well and is capable to capture a wide spectrum of attacks [13].

A network flow (*NetFlow*) is defined as an unidirectional sequence of packets with some common properties that pass through a network device, i. e., IP addresses, protocol and ports [4]. These flow statistics were originally generated by routers and switches for accounting and management purposes only. Nowadays, there are many network devices (including stand-alone probes) exporting NetFlow for the purposes of network behavior analysis and the anomaly detection. The detection is feasible even in 10 Gbps+ networks without any packet loss using flow-based approach, because the flow exporting process inspects only packet headers, not the entire packet payload. In our experience of deploying and running many NetFlow probes at campus network, NetFlow monitoring is very usable and powerful tool.

In the following work, we are discussing possibilities to detect the botnet traffic by inspecting DNS network traffic stored in the NetFlow format. This work is based on the fact that each bot performs DNS queries in order to connect to the command and control (C&C) server or to download its update from the

botnet server. Existing methods distinguish legitimate DNS queries from the botnet queries using various types of metrics, e.g., the amount of DNS queries going outside the local network, the frequency of dynamic DNS usage, similar DNS behavior of host groups etc. A main problem is that these existing methods are using data and information contained in data payload and not present in the NetFlow statistics. Therefore, we will try to examine existing methods and their possibilities to apply them on the NetFlow data.

Especially, we will focus on the following questions:

- Is it possible to detect botnets from the NetFlow data with no knowledge about packet payload?
- Is it possible to detect botnets by monitoring group activities of DNS queries?
- Is there a relationship between DNS connections going to the outside (non-local) DNS servers and possible botnet infection?

In this report, we provide an overview of the DNS traffic generated by the botnets (Section 2), then we provide a description of existing botnet detection methods using DNS traffic (Section 2.2) and we list important features contained in DNS traffic suitable for botnet detection (Section 2.3). Following Section 3 discusses the NetFlow data and what information about DNS traffic it contains. An analysis of NetFlow data containing DNS traffic is presented in Section 4. The possible directions for future work are suggested in Section 5.

2 DNS Traffic Generated by the Botnets

Domain Name System (DNS) [15] associates various information with domain names assigned to each of the participating entities. Most importantly, it translates domain names meaningful to humans into the numerical identifiers associated with networking equipment for the purpose of locating and addressing these devices worldwide.

DNS is one of the core services of current Internet. It is used not only for obvious benign purposes, but also for malicious use. As example we can see its usage in case of botnet command and control servers (C&C), phishing sites or download sites with malicious code.

If we will take a look to the existing botnet infrastructures more in detail, we will see a strong need for management of large number of bots. These bots need to communicate with C&C centers to receive commands and to pass harvested information. The IP addresses of C&C servers cannot be hard-coded to the binary codes of bots - in such case, it would be easy to take down a particular botnet by blocking this particular IP address of command server.

Therefore, by using DNS services, the attackers are able to change IP addresses of C&C servers with no need to modify bot codes (where is hard-coded only URL address of C&C server). They have also a possibility to hide malicious servers behind the proxy services using FastFlux technology [8].

2.1 Specific Aspects of Botnets DNS Queries

If we compare DNS queries generated by benign hosts and by malicious sources, there is enough differences to be able to distinguish between malicious and

regular DNS query, as described in [2]. We can find various methods described in literature to differentiate regular and botnet DNS queries.

Specific behavior of botnets is implied by the following characteristics [1]:

- **Botnet structure one to many** – the relationship between botmaster and bots is usually one C&C server to many bots and therefore we can detect similar group behavior of all bots in the network traffic.
- **Botnet synchronization** – as the C&C server issues the command for bots, they communicate in the same time and also perform attacks in the same time. Therefore we are able to detect the increased amount of traffic related to the bot group compared to the traffic generated by the benign hosts.
- **Bots response time** – as the bots receive a command from the botmaster, they perform requested activities with a constant response time compared to a wide variety of response times in the case of legitimate host. Therefore we can measure the response time to discover bots presence in the network.

This specific behavior is present also in the DNS traffic generated by the botnets. E.g., the move of C&C server to a new location, when the old one is blocked, generates massive group DNS queries to find the location of the new C&C server. In these requests, botnets differentiate by the fixed group size generating DNS queries and also by the activity of botnet groups appears immediately compared to continuous and random activity of legitimate hosts [3].

2.2 Existing Detection Methods

The existence of botnets and botnets wide spread lead to many research studies focused on botnets behavior, detection, classification, etc. There are also some works focusing on the use of DNS queries for analysis and revealing botnets in ordinary traffic. In the work of Dagon et al. [7] were identified key metrics for measuring the utility of a botnets and in the following work [6], the authors analyzed canonical DNS request rate and compared DNS density. Similar approach was presented in the work of Kristoff et al. [9].

Techniques and heuristics for detecting DNS blacklist (DNSBL) reconnaissance activity, where botmasters perform lookups against the DNSBL to determine whether their spamming bots have been blacklisted, is suggested in [12]. Anomaly based botnet detection mechanism focusing on a group activity in DNS queries simultaneously sent by distributed bots is presented in [3].

Data mining approach was conducted in [14]. Approach identifying abnormal domain names issued by the malicious botnets and also analyzing DNS traffic requested by group of hosts is presented in [10]. Finally, the complex work by Bilge et al. [2] uses passive DNS analysis, examines a wide set of DNS traffic features and incorporates machine learning techniques.

2.3 DNS Query/Answer Features

By studying the DNS behavior of known malicious and benign domains [2], we are able to identify distinguishable generic features that are able to define the

maliciousness of a given domain. From the DNS queries and DNS answers we can retrieve information like the name of the domain queried, the time the query is issued, the duration the answer is required to be cached (i.e., TTL) and the list of IP addresses associated with the queried domain. From this information, the authors of [2] identified 15 different features suitable in the detection of malicious domains.

2.3.1 Time-Based Features

These set of features is based on the time, when the request was made. The time itself isn't very useful by itself, however, when we analyze many requests to a particular domain over time, patterns indicating malicious domain may emerge. The authors propose to use following set of features:

- **Short lived domain** – a domain, which suddenly appears in the global scope time series and disappears after a short period of activity. If a domain is benign, even if is is not very popular, the number of queries should exceed the threshold at least several times during the monitoring period.
- **Daily similarity** – this feature checks if there are domains that show daily similarities in their request count change over time.
- **Repeating patterns** – this feature aims to detect regularly repeating patterns.
- **Access ratio** – this feature checks whether the domain is generally in an idle state or is accessed continuously.

2.3.2 DNS Answer-Based Features

The DNS answer from the DNS server can contain several DNS A records. In such cases, the DNS server cycles through the different IP addresses in a round robin fashion. This technique is useful for load balancing. Attackers typically use domains that map to multiple IP addresses, and IPs might be shared across different domains.

- **Number of distinct IP addresses** – the number of IP addresses resolved for a given domain during defined time window.
- **Number of distinct countries**.
- **Number of domains sharing the IP with**.
- **Reverse DNS query results** – number of reverse DNS queries of the returned IP addresses.

2.3.3 TTL Value-Based Features

Time To Live (TTL) specifies, how long the corresponding response for a domain name should be cached. Ordinary values are set between 1 and 5 days. Setting lower TTL values is useful for the attackers. Using this approach, malicious systems achieve higher availability and become more resistant against DNS

blacklisting and take downs. Typical example are Fast-Flux Service Networks (FFSN) [8].

- **Average TTL** – simple TTL average, used in various detection methods.
- **Standard deviation of TTL**.
- **Number of distinct TTL values**.
- **Number of TTL changes**.
- **Percentage usage of specific TTL ranges** – malicious traffic tends to set their TTL values to lower values.

2.3.4 Domain Name-Based Features

The main difference between regular services and malicious servers is that regular services try to choose domain names that can be easily remembered by users. In contrast, attackers are not concerned that their domain names are easy to remember.

1. **Percentage of numerical characters in domain name**.
2. **Percentage of the length of the longest meaningful substring**.

3 Flow Data and Its Content

Presented existing methods for using DNS traffic for detection of botnets are based on the assumption that we have complete payload data, which we can use for the features extraction. From the payload, we are able to mine out the domain name which was requested, TTL values, number of distinct IP addresses in the answer, etc. Contrary, in the NetFlow data, we have only limited amount of information about the DNS query and DNS answer. Namely, we have the following items:

- IP address of the host sending DNS query.
- IP address of the DNS server.
- Time of the DNS request.
- Time of the DNS answer.
- Size (in number of packets/bytes) of the DNS request.
- Size (in number of packets/bytes) of the DNS answer.
- Source port of DNS query and destination port of DNS query.

Having this limited set of items to use for botnet detection, compared to the various features contained in DNS query/answer payload, we are not able to use presented DNS detection methods and we have to focus on the other possibilities, how to detect botnets.

In the following, we will consider following features, which can provide us at least some information about possible maliciousness of particular host in the network.

- Usage of local DNS servers versus public DNS servers – ordinary hosts in local network are using local DNS servers to perform DNS queries. In the case of infected computers, they can use their own DNS servers or free DNS services (OpenDNS/FreeDNS). This behavior is possible also in the case of regular benign hosts, but the large amount of DNS queries against outside DNS servers indicates to possible botnet infection.

- Time of the DNS query – group of DNS requests from a group of hosts performed in the same time or in the small time window can indicate a possible command issued from C&C server and its move or update of botnet control servers. Therefore the bots perform DNS request to update addresses of botnet control servers.

As we can see, the amount of information suitable for the detection of botnets is very limited in the case of DNS traffic and NetFlow data. In the following, we analyzed a large sample of real traffic with the purpose to evaluate these two possible features and to determine, if they can be used for reliable detection of botnets.

Crucial is a place, where is the NetFlow probe deployed. There are big differences in the obtained NetFlow data in the case of NetFlow probe deployed on the backbone link compared to the NetFlow probe deployed inside the local network. The type of DNS traffic data differs significantly:

- NetFlow probe deployed outside the local network – in this case, the NetFlow data contains only the communication between DNS servers and part of DNS queries going from local network to the public DNS servers.

- NetFlow probe deployed inside the local network – in this case, we are able to inspect both DNS queries against the local DNS servers and also against the public DNS servers. Therefore, we have better data for the detection of possible botnet infection.

Another way, how to improve botnet detection, is to employ additional tools for extracting crucial information from the packet payload and add them to the flow data. Current NetFlow format does not support such extension of flow data, but in the case of IPFIX format [5], we are able to add such information to each flow and consequently, to deploy previously presented botnet detection methods using more traffic features.

4 DNS Traffic Analysis

To evaluate DNS features available in the NetFlow data, we captured one week traffic from real network and performed analysis of the DNS traffic contained inside this data. Our analysis was focused on the amount of DNS traffic compared to the total traffic, a ratio between UDP DNS queries and TCP DNS queries, how many DNS queries were performed against public DNS servers and if there are some detectable group activities in the DNS queries, revealing possible infected host in the monitored network.

The results of the analysis are summarized in the following table and figures. Table 1 provides overview of analyzed NetFlow data and amount of particular

types of DNS traffic. One week traffic was captured from the link with the average load 1,4 Gbps transferring 13,2k flows per second in average.

Data Type	Flows	Packets	Bytes
All traffic	8.0G	142.6G	104.9TB
DNS traffic	491.8M	559.2M	77.0G
DNS UDP traffic	490.2M	554.4M	76.7G
DNS TCP traffic	1.6M	4.8M	295M
Public DNS traffic	45.7M	60.4M	4.2G

Table 1: Overview of the analyzed DNS NetFlow data.

In the overall traffic volume, the DNS queries and replies represent 6.1% of data. If we will focus on the amount of TCP DNS traffic compared to UDP TCP queries, we can see that TCP protocol is used very rarely. Usually it is used when the response data size exceeds 512 bytes, or for tasks such as zone transfers. But as we can see from the Table 1, TCP DNS traffic represents only 0.3% from overall DNS traffic.

As we discussed in Section 3, the increased amount of DNS queries against public DNS services can indicate possible botnet infection. In the analyzed NetFlow data, 9.3% from all DNS queries were targeted to the public DNS services. There were no signs that these queries belong to the infected hosts. The overview of DNS queries against the public DNS services is illustrated on Figure 1.

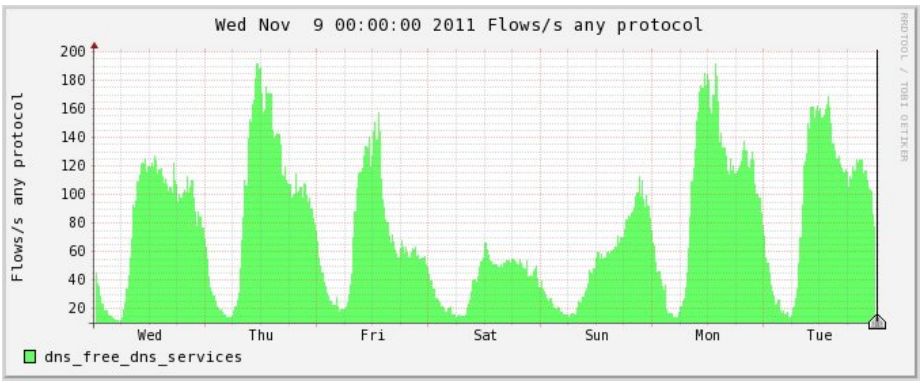

Figure 1: DNS queries to public DNS services.

The start time of DNS queries can reveal also botnet presence in the monitored network (see Section 3). Therefore we performed the analysis of 5 minutes windows and estimated the number of queries performed in each time window. Again, there were no remarkable events. The amount of DNS queries (i.e., the amount of DNS flows) is illustrated at Figure 2. In the case of massive network infection by the botnet, there will be remarkable DNS query events, but in ordinary traffic, we are not able to use this method to detect botnets using NetFlow data only.

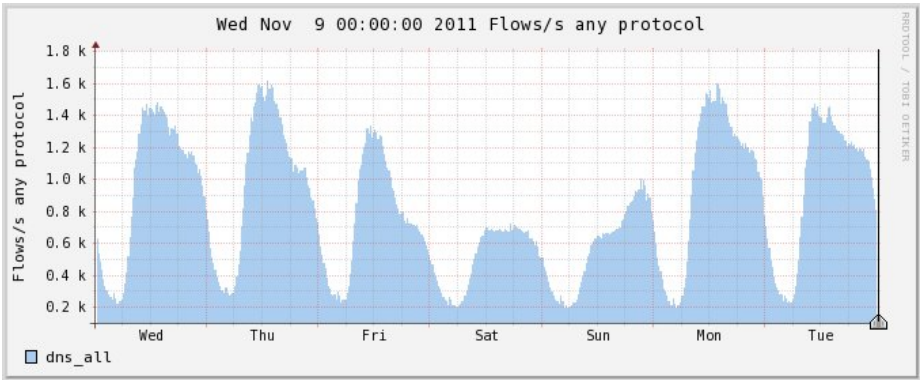

Figure 2: Amount of DNS queries during one week period.

5 Conclusion

In this internal deliverable, we focused on the possibility to use DNS flow traffic for purposes of botnet detection. In the first part of deliverable, we provided the description of the botnet generated traffic, we introduced existing methods for botnet detection using DNS traffic and we mentioned the set of features used for malicious traffic detection. The following part discussed the flow data itself, which have very limited amount of information about DNS queries compared to the full packet payload. The analysis of one week traffic from real network demonstrated the lack of crucial information contained inside the NetFlow data.

One feature, which we are able to monitor with current NetFlow data, is the amount of DNS queries generated from local network against public DNS servers. The aberrant amount of DNS queries against these DNS servers may indicate possible infection of local host by malware or botnet.

To conclude, we have to state that using NetFlow data solely, for the purposes of botnet detection, is not possible. There are several ways how to solve this problem. The most promising approach is to extract the important information from packet payload (queried domain names, their TTLs, IP addresses, etc.), use newer IPFIX export format and add the extracted information to each exported flow. In such situation, we are able to use existing botnet detection methods analyzing DNS traffic with advantage.

6 References

[1] Mitsuaki Akiyama, Takanori Kawamoto, Masayoshi Shimamura, Teruaki Yokoyama, Youki Kadobayashi, Suguru Yamaguchi. *A Proposal of Metrics for Botnet Detection Based on Its Cooperative Behavior*. Applications and the Internet Workshops, 2007. SAINT Workshops 2007. International Symposium on, s 82, jan. 2007.

[2] Leyla Bilge, Engin Kirda, Christopher Kruegel, Marco Balduzzi. *EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis*. NDSS, 2011.

[3] Hyunsang Choi, Hanwoo Lee, Heejo Lee, Hyogon Kim. *Botnet Detection by Monitoring Group Activities in DNS Traffic.* Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference on, s. 715 –720, oct. 2007.

[4] B. Claise. Cisco Systems NetFlow Services Export Version 9. RFC 3954 (Informational), October 2004. URL `http://www.ietf.org/rfc/rfc3954.txt`.

[5] B. Claise. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. RFC 5101 (Proposed Standard), January 2008. URL `http://www.ietf.org/rfc/rfc5101.txt`.

[6] David Dagon. *Botnet Detection and Response The Network is the Infection.* OARC Workshop 2005, 25, 2005. URL `http://madchat.fr/vxdevl/papers/avers/oarc0507-Dagon.pdf`.

[7] David Dagon, Cliff Changchun Zou, Wenke Lee. *Modeling Botnet Propagation Using Time Zones.* NDSS, 2006.

[8] Thorsten Holz, Christian Gorecki, Konrad Rieck, Felix C. Freiling. *Measuring and Detecting Fast-Flux Service Networks.* NDSS, 2008.

[9] J. Kristoff. *Botnets.* NANOG 32, 2004.

[10] Ahmed M. Manasrah, Awsan Hasan, Omar Amer Abouabdalla, Sureswaran Ramadass. *Detecting Botnet Activities Based on Abnormal DNS traffic.* CoRR, abs/0911.0487, 2009.

[11] Namestnikov, Y. The economics of Botnets, 2009. URL `http://www.viruslist.com/analysis?pubid=204792068`.

[12] Anirudh Ramachandran, Nick Feamster, David Dagon. *Revealing botnet membership using DNSBL counter-intelligence.* Proceedings of the 2nd conference on Steps to Reducing Unwanted Traffic on the Internet - Volume 2, s. 8–8, Berkeley, CA, USA, 2006. USENIX Association. URL `http://dl.acm.org/citation.cfm?id=1251296.1251304`.

[13] Anna Sperotto, Gregor Schaffrath, Ramin Sadre, Cristian Morariu, Aiko Pras, Burkhard Stiller. *An Overview of IP Flow-Based Intrusion Detection.* IEEE Communications Surveys & Tutorials, 12(3):343–356, 2010. URL `http://doc.utwente.nl/72752/`.

[14] Hao Tu, Zhi-tang Li, Bin Liu. *Detecting Botnets by Analyzing DNS Traffic.* In Christopher Yang, Daniel Zeng, Michael Chau, Kuiyu Chang, Qing Yang, Xueqi Cheng, Jue Wang, Fei-Yue Wang, Hsinchun Chen, editors, *Intelligence and Security Informatics*, volume 4430 of *Lecture Notes in Computer Science*, s. 323–324. Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-71548-1.

[15] Wikipedia. Domain name system — wikipedia, the free encyclopedia, 2011. URL `http://en.wikipedia.org/w/index.php?title=Domain_Name_System`.