Active DNS Collection Challenges (Abstract)

Athanasios Kountouras*, Panagiotis Kintis*, David Dagon*, and Manos Antonakakis*

*Georgia Institute of Technology,
{kountouras,kintis,manos}@gatech.edu, dagon@sudo.sh

I. Introduction

Most modern cyber crime leverages the Domain Name System (DNS) to attain high levels of network agility and make detection of Internet abuse challenging. The majority of malware, which represent a key component of illicit Internet operations, are programmed to locate the IP address of their command-and-control (C&C) server through DNS lookups. To make the malicious infrastructure both agile and resilient, malware authors often use sophisticated communication methods that utilize DNS for their campaigns.

To effectively combat Internet abuse, the security community needs access to freely available and open datasets. Such datasets will enable the development of new algorithms that can enable the early detection, tracking, and overall lifetime of modern Internet threats. To that end, we have created a system, Thales, that actively queries and collects records for massive amounts of domain names from various seeds. These seeds are collected from multiple public sources and, therefore, free of privacy concerns. The system supports the Active DNS Project [1], which opens DNS dataset that can enable both repeatable DNS research and more efficient threat research for the security community.

Infrastructure: Our existing project is composed of a traffic generation component, a data collection and processing component and finally a domain seed. Our traffic generation is based on a distributed system of a client and recursive server pair, with each pair contained in a Linux Container. The traffic generation component is responsible for querying the domains found in the domain seed. The client consumes the domain seed, and sends queries for the domains to the recursive. The traffic is collected by a separate machine through a span port that mirrors all the generated traffic. A Hadoop cluster then handles the parsing of the network captures and the conversion to a usable and efficient format for research. Thales is currently generating 1.7 TB compressed DNS traffic. After a daily deduplication process we are left with 85GB records in AVRO format.

Measurements: Utilizing the current infrastructure we are able to resolve 290 million QNAMES every day, and collect 1.2 billion of resource records after the daily deduplication. The system is continuously running and queries every domain in the domain seed at least once daily. We actively query for A, AAAA, SOA, MX and TXT QTYPES with A and NS records being the most populous in our collected data. You can find some simple statistics about the ongoing daily operation of our project in https://www.activednsproject.org/statistics.html

II. Growing The Project

The project was fortunate enough to receive a warm and enthusiastic reaction from the security community. A lot of the academic and industry research labs have subscribed to our daily data update and are actively trying to find ways to contribute back to the project. The growth of this project in size and in the value of data that we collect and share with the scientific community is a key goal for our lab. Up to this point the simplest way to help the Active DNS project is to contribute domains that can be added to our daily querying seed.

Perhaps one of our most critical next steps is to increase the geographic distribution of Thales' querying infrastructure. The simplest way to achieve that would be to utilize the power of cloud-based infrastructure. However, the monetary costs associated with such an effort makes it not feasible according to current lab resources. Thus, the simplest way that researchers could help this project is by providing hosting infrastructure for data sharing, or start up their own collection infrastructure, or by sponsoring the costs for creating more Thales instances in the cloud.

The challenges we face in our effort to expand the project are also technical. The decisions we will make in the way we expand, will shape the future of the project and the quality of the dataset. Because this is a community project we would like to have a consistent way to grow. Such growth will enable people to build similar systems, and potentially share their data back to the community. At the same time, we want to keep the shared dataset consistent for both repeatability and ease of use.

Summarizing the challenges we can identify and the questions that we want to raise are: (1) How to replicate the system on outside contributor's networks. (2) What would be the ideal method to curate the domain seed to increase the value of the generated dataset for more diverse applications. (3) What is the best way to share DNS data more efficiently and effectively. (4) Any interesting QTYPES that we should consider for collection (i.e., DNSSEC QTYPEs). (5) Ways to increase the consistency between different collection infrastructures. (5) What is a good policy for sharing data generated by third parties. (6) Could we explore low bandwidth options and less resource intense solutions for people that want to contribute but have limited resources. (7) Potential new research ideas that can be investigated with the current or a modified dataset. (8) Finally what other subsets of the existing dataset would have value for researchers working with DNS data.

REFERENCES

[1] A. Kountouras, P. Kintis, C. Lever, Y. Chen, Y. Nadji, D. Dagon, M. Antonakakis, and R. Joffe, "Enabling network security through active DNS datasets," in *RAID*, ser. Lecture Notes in Computer Science, vol. 9854. Springer, 2016, pp. 188–208.