CyberGreen

Data Validation Report

June 2017

Executive Summary

To build its internet health metrics, CyberGreen collects data from third parties using different methodologies and tools. In order to improve the collection and analysis of this data, CyberGreen cross-references the data against each other and over time. This report discusses initial findings about the data and how CyberGreen will use these findings to launch its own scanning capability.

Our primary observations are that there is a critical need to cross-reference multiple data sources, as different techniques provide different results. We also note that many of the data sets are highly volatile, with a large replacement rate for IP addresses in consecutive scans.

Introduction

In order to develop effective metrics for internet-wide risks, CyberGreen must collect data from multiple diverse sources. In the initial phases of this project, CyberGreen has focused on collecting sources to estimate DDoS risk. These sources may include, but are not limited to, information on DDoS reflectors (e.g., DNS, NTP and other UDP-based servers which will send a packet in response to any request) and IoT devices (subject to mass takeovers as exemplified by the Mirai botnet).

This report is an initial survey of the third party data CyberGreen currently uses. The intent is to use this data to provide an overview of what information is currently available and use that data to inform CyberGreen's own data collection and scanning. Based on the data collected, we observe that the problem of *accurately* evaluating the data collected by scans is nontrivial -- multiple challenges exist to the validity of the scanned data. The process of developing a gold standard for data collection, one which can be used as a reliable intelligence feed, is an ongoing process of discovery and remediation, as the data collected from scans and intelligence informs us of errors which we can rectify to continuously improve data collection.

The solutions discussed in this report are framed as design requirements for CyberGreen's proposed scanning capability. We note, however, that the most important observation we have made from this data is the need for diverse data collection; CyberGreen will continue to collect multiple sources and add to our collection over time, with the intent of cross-referencing and validating the data to produce more effective results.

Issues of Validity in Scan Data

We will now discuss our efforts to evaluate the quality of the data provided; to do so, we have drawn on the concept of *validity.* The *validity* of an argument refers to the strength of the conclusions one can generate from an inference[1]. An analyst cannot demonstrate that an argument is valid; instead, they must address *challenges* to the validity, such as the possibility of external interference in data collection. Researchers have identified different classes of validity with their own challenges; for our purposes, we focus on one particular class: *internal validity*. The internal validity of an argument refers to the strength of the argument's assertion of cause and effect.

The internal validity of the scan data refers to the strength of the argument that the *response* to a scan (either that the target did or didn't send a packet back) is an indicator of the presence or absence of a relevant host for a reasonable interval. We can more formally state this as:

1. [True Positive] If an IP address responds in a short interval to a scan packet sent to it, then for some reasonable time, that IP address is a host relevant to us (i.e., it contains DNS, SNMP, etc.)
2. [True Negative] If an IP address *does not* respond in a short interval to a scan packet sent to it, then for some reasonable time, that IP address is not a host relevant to us.

There are multiple challenges to the validity of these inferences, in particular:

- [False Positive]. IP addresses may be transient. DHCP is standard practice for many networks, which means that many hosts may be moved to new addresses in a short interval.
- [False Positive]. A system, particularly network hardware, may opt to respond to every packet which it receives.
- [False Negative]. Firewall rules between the target and the scanner may drop the incoming packet, resulting in the scan never reaching the target.
- [False Negative]. A route or network may be temporarily disrupted, such as by a DDoS.
- [False Negative]. UDP scanning packets must implement enough of the targeted service to generate a response. They may not correctly do so for all implementations.

To compensate for these challenges to internal validity, we will need to scan diversely, and repeatedly. Diverse scanning will require scanning from multiple locations over time, as well as scanning with different forms of the same service packets; this will enable us to compensate for accidents of routing which may block traffic, as well as accidents in the implementations of the various services we are profiling.

[1] W. Shadish, T. Cook, and D. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Stamford, CT: Cengage Learning, 2001.

Of particular importance for this effort is figuring out the minimum sizes and intervals necessary for meaningful inference. Given DHCP and dynamic allocation individual IP addresses can be an imprecise indicator. In dynamically allocated networks, it may be more effective to consider constructs which describe the network in aggregate. In the case of time, the question of the lifetime of a scan result is critical -- our scans must repeat within a short enough interval that we can trust the results, but not so often that we risk aggressive backlash from the organizations which we are scanning.

Internal validity is one of four major classes of validity, the others being *external* (the generalizability of the observations), *construct* (the models used to describe our analyses), and *statistical* (the statistical techniques applied to constructs). as we acquire a better understanding of the data, we will address further challenges to the validity -- in particular, the generalizability of the results, and the lifetime of transient addresses.

Overview of Data Sources

For this initial report, we have examined five datasets, which we refer to as *OpenResolver UPNP, SNMP, NTP and DNS,* and *CenSys UPNP*. The OpenResolver sets refer to UPNP (or SSDP), SNMP, NTP and DNS data collected by a single host from the OpenResolver project[2]. Each scan is conducted weekly, on a separate day and then converted into csv data by CyberGreen's ETL process. The CenSys UPNP data consists of output from the CenSys scans project[3], an internet-wide scanning supported by the University of Michigan; it is smaller, varies more, and is considered supplementary.

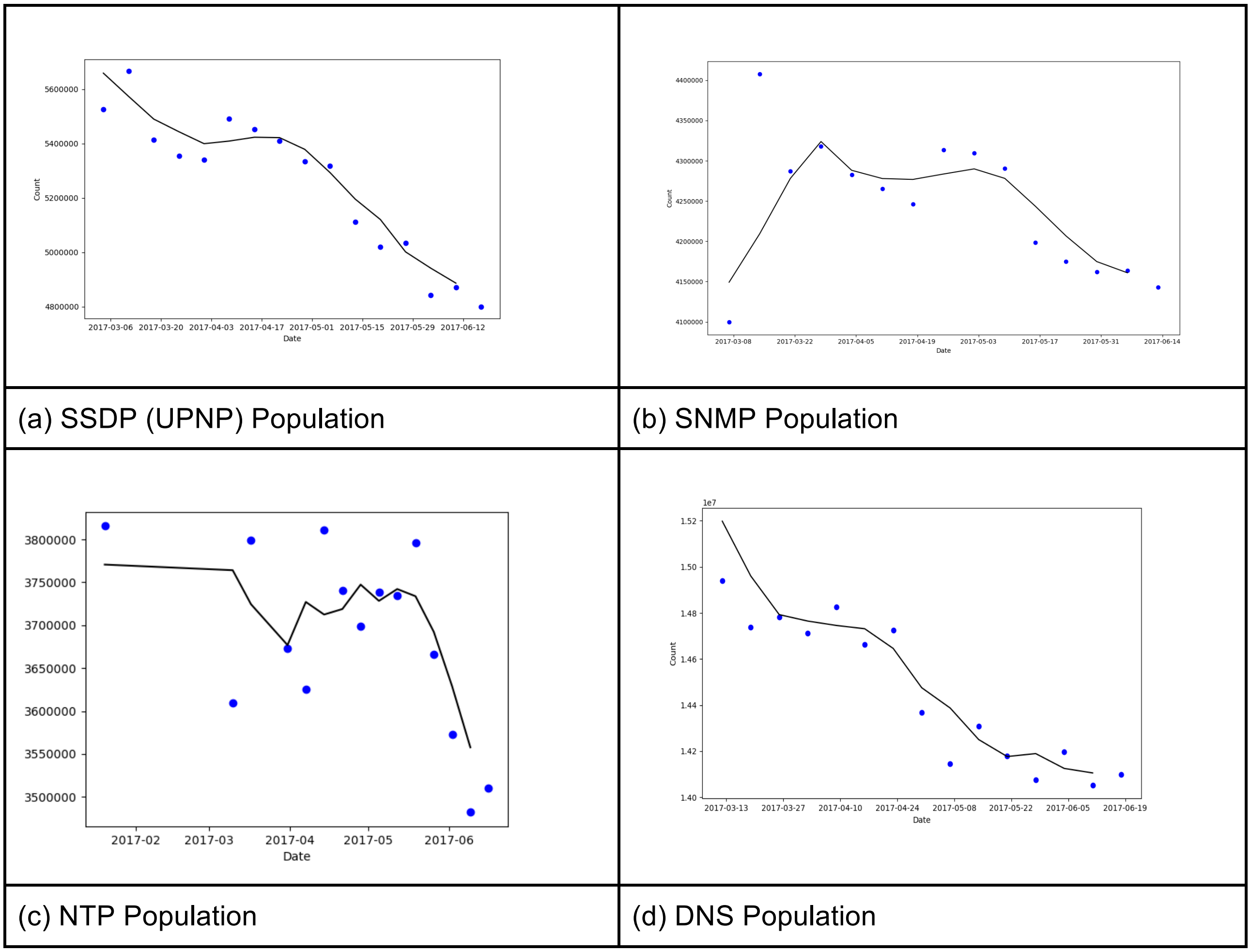
Analyses of Data Sources

In this section, we analyze the collection of the data sources in order to address the problem of internal validity. The issues discussed in this section focus primarily on the completeness of the data covered, and how the data from these sources change over time. This section discusses three separate topics -- the aggregate change in population over time, the volatility of the addresses observed, and the issue of the completeness of the scans. We discuss challenges in the existing data, and how to compensate for them in future work.

Aggregate Populations

Figure 1 shows the aggregate populations of hosts responding to the four scans in 2Q 2017, approximately March-June 2017. As this figure shows, the populations decrease consistently over time, approximately a 1 percent relative drop week over week.

[2] http://openresolverproject.org
[3] http://censys.io



(a) SSDP (UPNP) Population

(b) SNMP Population

(c) NTP Population

(d) DNS Population

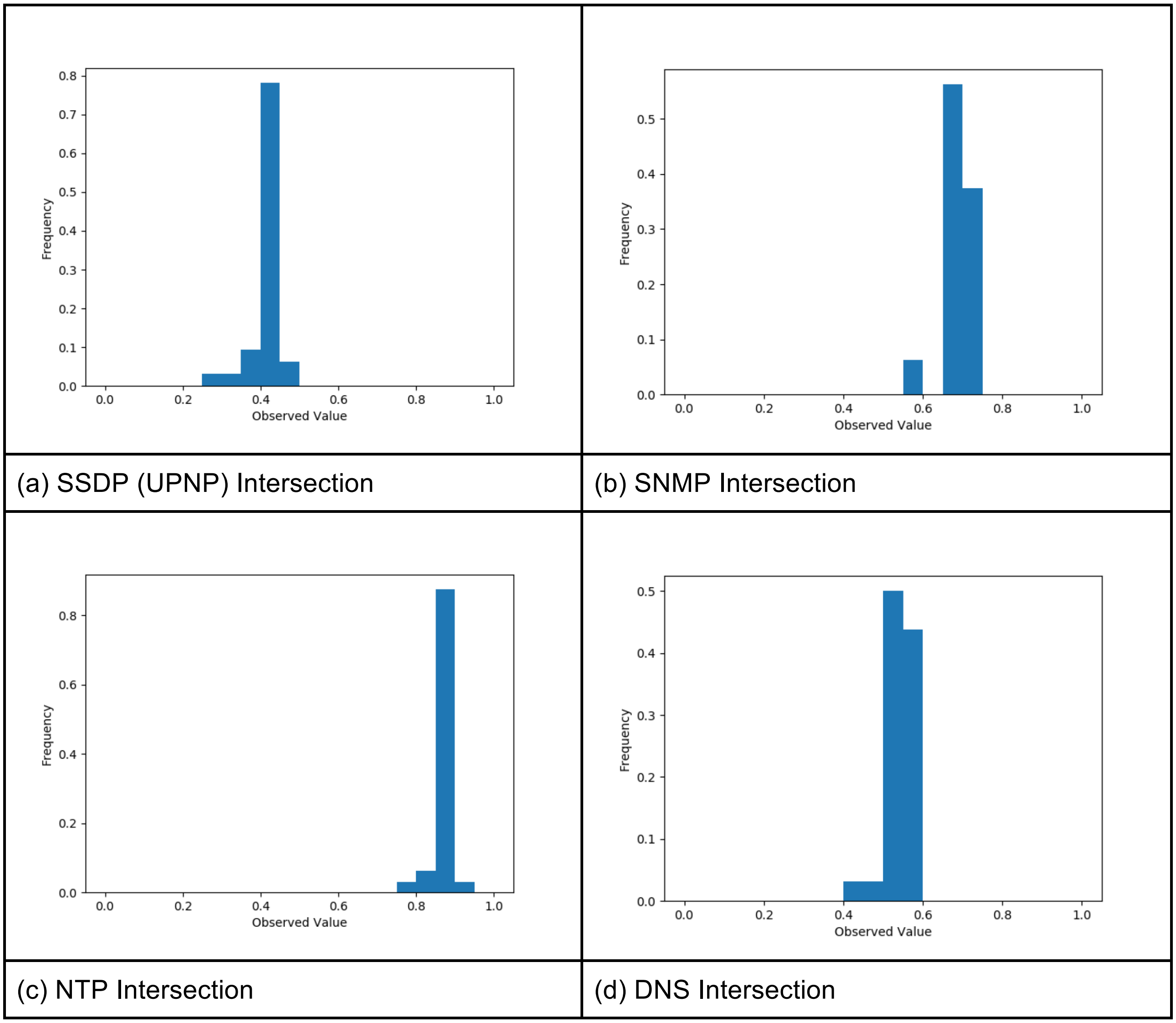
Figure 1: Change in Scanned Host Populations, 2Q 2017

In comparison to the OpenResolver data, the Censys UPNP dataset is smaller and more varied. The CenSys datasets contain between 400,000 and 1.46 million hosts per scan, with the numbers rising or falling by more than 50% each week. By way of comparison, the current OpenResolver data contains approximately 5 million discrete hosts. We are not, at this time, sure why the observed difference occurs, but expect to have better points for comparison by scanning with zmap, the same tool CenSys uses.

We hypothesize that the consistent drop in the OpenResolver population is due to an increasing number of targets blocking the OpenResolver scanner. The project is long-lived and a single host conducts the scanning. The most likely alternative hypothesis is that the targeted networks are patching their hosts, resulting in a smaller population. We suspect that this is unlikely, as the decrease is consistent across the different OpenResolver sets, and the same behavior is not observed in the CenSys data. We expect to test this hypothesis by scanning from multiple diverse locations and looking for evidence of progressive blocking over time.

Address Volatility

By *volatility* we refer to the likelihood an observed address will appear in multiple consecutive scans. Understanding the volatility of the observed addresses is critical for establishing a valuable lifetime for the scan data. Figure 2 shows comparative histograms for the probability that an address will appear in two consecutive OpenResolver datasets. That is, the probability that if an address appeared in week 1, that it also appears in week 2.



(a) SSDP (UPNP) Intersection

(b) SNMP Intersection

(c) NTP Intersection

(d) DNS Intersection

Figure 2: Comparative Volatility of OpenResolver Datasets

As Figure 2 shows, the address volatility of the individual datasets is service-dependent. A rough ordering is that NTP is the least volatile, with approximately 85% of the addresses in one week appearing in the next. This followed by SNMP (approximately 65% stable), DNS (50%) and then UPNP around 40%. We hypothesize that this volatility is due to a large number of cheap embedded appliances such as DSL modems or wireless routers.

The volatility of the addresses raises an important question about the value of individual IP addresses as a construct. If it is the case that many of the individual addresses are assigned by DHCP then individual IPs within a DHCP netblock may be less valuable than an aggregate population count. In future work, we propose to divide the population into a consistent set of servers (i.e., ones that appear in multiple scans), and a transient set. This set may be discovered through identifying DHCP netblocks, scanning hosts for other indicators, or by selectively rescanning hosts.

Comparing Scanning Sets

The OpenResolver project and CenSys both scan hosts for UPNP/SSDP activity, enabling us to cross-correlate the two sets and determine how complete the coverage is. To do this, we used a rough estimate based on the Lincoln-Petersen estimate from capture-recapture analysis. Given two sets collected by different methods, A and B, the Lincoln-Petersen estimate of the total population is:

$$P=\frac{|A||B|}{|A \cap B|}$$

We would expect that if the two collection methods were producing identical results, then P=|A|=|B|. This is a rough estimate that will provide a mechanism for measuring the completeness of our scan efforts in future work. The better, more consistent and more complete our estimation techniques are, the smaller the difference between P and |A|.

Calculating P on sets scanned close in time yields an estimated P of approximately 19 million, over 3 times larger than the OpenResolver data for the same period. We note that these results are rough -- OpenResolver and CenSys use different methods to collect their scan data, and the different populations imply that blocking, transience and routing errors may result in incomplete data in either case. However, the fact that these ostensibly common data sets have such a low common population raises a strong possibility that all the scans are undercounting the population of vulnerable hosts.

Conclusions and Future Work

In this work, we have conducted an initial study of the data collected by OpenResolver and Censys in order to determine the quality of the data and provide guidance for future work. Based on this work, we have identified multiple challenges to the internal validity of this data, and proposed solutions. As CyberGreen develops its own scanning capability, these challenges will inform the design, implementation and operation of that capability.

By far, the most important observation about this data is that *no single scan is sufficient*. Geolocation, population volatility and false positives affect the results from any single scan, whether this is multiple scans from the same source over time or scans collected from multiple sources.

Developing mechanisms for managing address volatility is particularly critical for determining how to operationally scan networks. The observed volatility in the OpenResolver datasets suggests that we need to either scan networks more frequently than once a week (potentially triggering additional blocks) or develop constructs to compensate for this volatility. We expect to implement a hybrid solution by partitioning the network into volatile and nonvolatile subnetworks, and targeting scans appropriately.

Based on the population volatility, we are concerned about whether individual IP addresses are a meaningful element. If DHCP is affecting addresses as heavily as observed, it is likely that a significant fraction of the scanned population has already moved by the time the scan completes. As we develop our risk models, we will have to compensate for this volatility, possibly by partitioning the internet into static and dynamic segments and using different values or risk indicators for each.

We envision the CyberGreen scanning capability as a controlled, clearly-defined set of scan data which we can use to compare against other data sets. Using estimators such as the Lincoln-Petersen population estimator discussed above, we can develop a metric for the completeness of our coverage. In future releases, we will publish this as an estimate for how close we are to achieving a complete estimate of hostile agents.