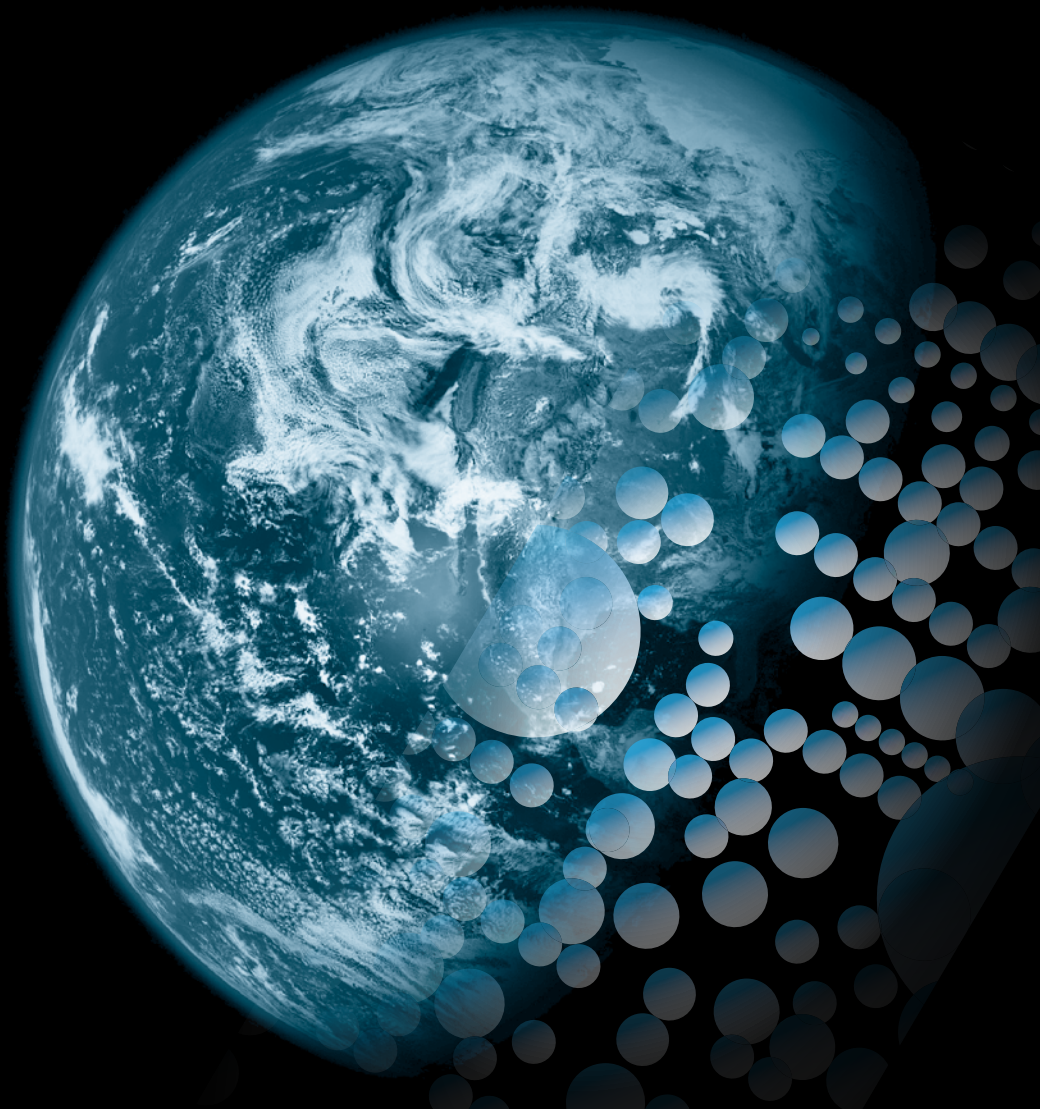
Ed Stoner

Finding Malicious Activity in Bulk DNS Data



Extracted from the 2009 CERT® Research Annual Report

Finding Malicious Activity in Bulk DNS Data

Abstract

The Domain Name System is a vital component of the Internet, and nearly every transaction on the Internet uses it. It contains a wealth of Network Situational Awareness information that can be used to discover malicious traffic. This report describes specific techniques to detect certain types of malicious traffic. These techniques have been developed through analyzing a large amount of DNS traffic data. CERT has developed specific tools that apply these techniques in an ongoing way. Future research will include enhancing the developed tools, developing new techniques and tools to work with known malicious patterns, and discovering new malicious patterns.

Problem Addressed

The Domain Name System (DNS), which maps names to IP addresses, is a vital component of the Internet. Nearly every transaction on the Internet begins by making a DNS query. This is true for both benign and malicious activity. Having access to large amount of DNS queries allows us to look for patterns in Internet transactions. In certain cases, the patterns for malicious activity can be distinguished and identified.

Research Approach

The Security Information Exchange (SIE) is a framework for information sharing run by the Internet Systems Consortium (ISC). It gives researchers access to a large amount of DNS messages. CERT has a server setup at SIE that captures information for about 400 million DNS messages per day. DNS messages contain either a question about a host name (i.e. what is the IP address of www.google.com?) or an answer to a question. The DNS messages that are streamed by SIE are all answer messages that are marked as authoritative and do not contain errors. These messages are streamed to the CERT server and stored in the ncap file format developed by ISC specifically for this type of application. The ISC provides a piece of software called ncaptool to send, capture, and process these streams of DNS messages. With ncaptool and additional software developed at CERT, we are able to analyze the stored DNS messages for particular patterns of malicious activity.

Some of these patterns are the result of using the DNS infrastructure itself in a way that it was not originally designed for. A particular example of this is DNS tunneling. Other patterns of malicious activity are detectable because of certain constraints that the type of malicious activity places on how DNS can be used. An example of this is Fast Flux hosting.

DNS tunneling is a process where DNS messages are used to transport arbitrary data by encoding that data into the DNS messages themselves. Because of the very wide support and availability of the global DNS infrastructure, and because very few organizations block DNS traffic from individual clients to the Internet, this method can be very effective for bypassing security measures such are firewalls or ACLs.

A DNS tunneling implementation is detectable when it is created to encode arbitrary data, and is either used for two-way communication or data exfiltration. The reason for this is that in a DNS question (which is what a client would use to pass information outbound), the only place to encode information is in the host name. Per the DNS protocol specification (RFC 1035), the host name has only 63 allowable characters (all upper and lower case letters, digits 0 through 9, and hyphen). In order to encode arbitrary data and achieve reasonable bandwidth, implementations of this type of tunneling will use noticeably more unique characters than normal host name would have. Figure 1 shows the unique character counts for a host name used in DNS tunneling and for www.google.com.

host name	unique characters
08f0b06a25a5cf1f9df501bc39306 fbc6ff7875646817b4845c17da0.6. ewsxz.com	23
www.google.com	8

Figure 1: Unique character count of host names

It is also a feature of the DNS protocol that the question is contained in the answer message. So even though our data only has DNS answers, it is still possible to find the tunneling. This is done by iterating through a particular set of messages, extracting the host name in the question, and counting how many unique characters are in it. After some initial testing it was determined that 20 unique characters was a good starting point for finding DNS messages with encoded information.

In fast flux hosting, an attacker uses DNS to hide malicious sites behind an ever-changing network of compromised hosts. This pattern shows up in DNS records as an unusually large number of distinct IP addresses in answers returned for the query of a single domain name, with each answer having a very short period of validity or time to live (TTL), and with previously unseen IP addresses constantly emerging in queries over time. The answers given all point to a proxy network—a set of compromised machines that relay traffic to a central host or small set of hosts that the malicious party controls. This proxy network hides the real malicious site, making it difficult to track the site and to take it down.

To find Fast Flux hosting in our data, we first collect answers which give more than ten IP addresses for a given host name and have a TTL of 2000 seconds or less. Next, we count how many unique IP addresses are seen for each host name in the data we collected. Lastly, for host names with more than 25 unique IP addresses, we count how many different ASNs (Autonomous System Numbers, which map to Internet Service Providers) there are. If there are more than 20, then it is extremely likely that Fast Flux hosting is happening for the host name.

Answering with multiple IP addresses for a question about a host name is a long-standing legitimate practice to provide redundancy and high-availability. What makes Fast Flux different and detectable is that the hosts are compromised. Therefore, unlike a legitimate service, they are much more unreliable, so more hosts are needed, and they need to have a shorter TTL. Secondly, the hosts need to be on very diverse networks, or otherwise it would be easier to shut them off. Taking advantage of these characteristics yields a very successful algorithm.

In researching Fast Flux hosting, we noticed a few other characteristics that are prevalent in domains setup for malicious activity. One is that host names have a computer generated label. The other is that a top-level domain (like .com or .net) will appear in the middle of the name (www.somebank.com.badguy.tv). We can use these characteristics to sort out malicious domains from compromised domains in existing block lists (which can then be used to track down hosts that are visiting malicious domains).

Expected Benefits

By identifying malicious activity in DNS messages, we can begin to develop a more proactive monitoring approach. Rather than relying on hand-assembled lists of malicious activity, we can start automatically generating lists that have the potential to include zero-day and previously unnoticed attacks and attack vectors. This can provide for quicker incident response time and the ability to notice a wider range of incidents.

2009 Accomplishments

The tools to apply the Fast Flux detection algorithm described to large amounts of DNS data have been developed. A system has been put in place to find Fast Flux domains as well as hosts on certain networks that are connecting to those domains.

The tools to find DNS tunneling and exfiltration have been developed. Because there are so many examples of legitimate exfiltration, the techniques are very sensitive as to what network they are monitoring.

The tools to separate malicious domains from known lists of bad domains and find the IP addresses associated with those domains have been developed. These IP addresses can then be used by systems already in place that monitor activity by IP address.

2010 Plans

One challenge that came up in our research into finding DNS tunneling was determining which tunneling messages are malicious. Because of its effectiveness and wide support, DNS tunneling has been adopted by many organizations for providing Real-time Blackhole Lists and similar lookup services. Future work will include developing techniques to identify whether a tunnel is benign or malicious.

The current tools for distinguishing between malicious domains and benign domains are very crude, and simply check for a large number of unique characters and top-level domains that are in the middle of a name. Because of this they require existing lists of known bad domains. We are exploring new techniques to be able to detect malicious domains with more reliability and without pre-filtering. That work includes identifying better algorithms to determine if a label in a DNS host name is computer generated.

By continuing to develop and refine our toolset for analyzing DNS messages, we hope to identify additional malicious characteristics, and to provide additional tools to detect and report incidents. By combining the DNS data with other data sources we hope to detect new types of malicious activity.

References

[1] Tactical IT. “DNS Traffic Analysis.” http://tactical-it.com/2009/01/a-study-of-dns/

[2] Thorsten Holz, Christian Gorecki, Konrad Rieck, & Felix C. Freiling, “Measuring and Detecting Fast-Flux Service Networks,” 2008 ISOC Network & Distributed System Security Symposium.

[3] Jose Nazario, Thorsten Holz, “As the Net Churns: Fast-Flux Botnet Observations,” 3rd International Malicious and Unwanted Software (Malware 2008).