Correlating Domain Registrations and DNS First Activity in General and for Malware

Jonathan M. Spring, Leigh B. Metcalf, and Edward Stoner

Abstract—**From the date that a domain name is registered with a registrar, there should be a pattern in the amount of time it takes for that domain to be actively resolved on the Internet. We first attempt to describe that pattern in general terms by correlating data from registries for several top-level domains and a large passive DNS data source. This pattern is then used as a baseline for a comparison with the pattern of activity in domains that malicious software utilizes. While our quantitative results are not to be considered representative of the patterns exhibited by all types of malware, the malicious domains are found to have a significantly different pattern than the standard domains.**

Index Terms—**measurement studies, passive DNS, SIE, malware and the DNS.**

I. Introduction

DOMAIN names must be registered for use before they are accessed via the domain name system (DNS). Companies and individuals generally do business with registrars, who collect the necessary information and payment and then pass the new domain off to the appropriate registry. Both malicious and legitimate domains must be registered through this same process; however it is reasonable to suspect that there will be different patterns of behavior between the two types of domains.

To date, it seems that this correlation has not been made in general, and so a baseline pattern of behavior for the Internet in general must be established before any deviation from this norm could be measured. If the pattern discovered for malicious domains is sufficiently different from the average, then the hope is that this difference could become part of a method for detecting malicious domains before they do damage, rather than retroactively.

II. Procedure

A. Preparation

We collect data from a high-volume passive DNS source at the Security Information Exchange (SIE)[1]. This is a near-real-time feed of data collected from several high-volume DNS servers distributed throughout the Internet. The location of the sensors is not public, and so the bias introduced by the location of the sensors is not calculable. However there is some bias introduced, and this data cannot claim to capture all DNS data on the internet. Regardless, the SIE data represents the most comprehensive DNS data source available at this time. There is reason to believe the data is a sufficient sample size to move forward with. Reference [2] demonstrates that the SIE data provided visibility to resource records (RRs) for about 93% of the domain labels immediately under the .edu top-level domain (TLD) in a 2-week observation period.

Further attempts were made to measure the representativeness of the SIE data. The data is delivered in the nmsg format, and each message contains the IP address of the machine that sent the response. Counting the unique number of these IP addresses, and also to which autonomous system number (ASN) they belong, delivers some measure of the diversity of the responses captured by the SIE. Samples were taken for March 11 and 16, 2011 from channel 207; channel 207 reduces data volume by deduplicating exact copies of messages and incrementing a counter in the nmsg format. This does not affect the number of unique IP addresses observed. Over these two days, 1.56×10^9 nmsg messages were observed. The IPv4 addresses were simply extracted and counted. For the ASNs, the IP addresses were correlated using a comprehensive mapping of ASN to IP that CERT maintains internally and is updated daily. Correlation and storage utilized the SiLK toolset.

In generating the lists of newly active domains we reduce the SIE feed to a list of all of the unique RRs observed for a given day. These lists of RRs are then further processed in order to generate a list of all of the two-label domains (e.g. example.com) that were observed for the first time in our data collection on the current day. In order to provide a baseline for what was new each day, we calculated the new two-label domains every day starting June 1, 2010.

We collect zone data for the biz, com, info, mobi, and net top-level domains on a daily basis. From this data, we use a Bloom filter to create a list of the newly registered domains for that top-level domain (TLD) for a given day. For the month of October 2010, there were 2,783,497 domains registered in the TLDs that we have data for.

Manuscript received March 21, 2011.

Jonathan M. Spring, Leigh B. Metcalf, and Edward Stoner are members of the technical staff at the Software Engineering Institute, Carnegie Mellon University. 4500 Fifth Avenue, Pittsburgh, PA, 15213. {jspring , lbmetcalf , ers }@cert.org. 412-268-2000.

CERT is a branch of the Software Engineering Institute, a federally funded research and development center operated by Carnegie Mellon University.

Thirdly, we collect information about domain names that are related to malicious code. Within CERT, malicious code is collected and analyzed in the artifact catalogue [3]. Some of the results include the domain names that the malware attempts to resolve. This data makes up the corpus of the malware-related domains that we study. Data is available for both the third and fourth quarters of the 2010 calendar year. Of the 146,856 unique domains observed, 4,729 (3.2%) were found in the SIE data and could be correlated with zone registration files. The SIE database is indexed on domain names, and so is more efficient for these types of look ups. The SIE database zone file information reaches back to April 2010. However it only has TLD zone files and does not have zone files for every TLD, but instead seems to be similar to our zone file data in containing generic TLD data. This also excludes dynamic DNS services from evaluation. This is a potential bias in the evaluation, because we can only calculate the latency for those malware which happen to successfully lookup domains in these generic TLDs. The registration-request delay was calculated for each domain which had an entry.

B. Evaluation

Once these lists are generated, the list of newly registered domains for a given day is correlated with the lists of new second-level domains. This is done several months after the domains were registered to allow for sufficient chance that the domains are actively resolved on the Internet. For each domain on the registered list, the lists of new second-level domains are searched, and the day for which it is found, if any, is recorded. Currently, the domains registered on October 1 through October 31, 2010, have been evaluated for active lookups occurring between August 1, 2010 and January 24, 2010.

The dates for malware domain collection and evaluation are a superset of those for the general case. Malware reports are organized by when they were analyzed. Both the Q3 and Q4 domains have been evaluated by checking the entries in the SIE DNS database. The domains on the list for each quarter are searched for in the database and those with a zone time first seen and a DNS packet time first seen are evaluated. Additionally, the domains observed in Q4 that happen to have been registered in October 2010 are available for correlation with the general data from the zone files.

The data for the number of days that transpired between registration and lookup are then summarized with some statistics and counts.

III. RESULTS

Results for the diversity of the of the SIE data are encouraging, however not exhaustive. For March 11, 2011 788,998 unique IPv4 addresses were observed, which represented 181, or 70.7%, of the /8 CIDR blocks. 802,324 unique addresses were observed March 16, which covered 180, or 70.3%, of the /8 CIDR blocks. The union of the two sets of IPv4 addresses consisted of 875,972, also covering 181 /8 CIDR blocks.

The ASN coverage results for March 11 are 24,968 ASNs represented by the IPv4 addresses out of 36,551 that were routable that day, or 68.3%. On March 16 24,998 ASNs were represented, out of the 36,607 that were routable that day, or 68.3 %. The union of the sets of observed ASNs consists of 25,399 unique ASNs. The number of unique routable ASNs for the two days is 36,678. 69.2% of the routable ASNs were represented as the sender of at least 1 DNS response in the SIE data on these two days. These results are summarized in Table I.

TABLE I
TOWARDS EVALUATING THE REPRESENTATIVENESS OF THE SIE DATA

Sample Day	Observed unique IPs	/8 CIDR blocks (% of total)	Unique ASNs (% of total routable)
March 11, 2011	788,998	181 (70.7%)	24,968 (68.3%)
March 16, 2011	802,324	180 (70.3%)	24,998 (68.3%)
Total unique	875,972	181 (70.7%)	25,399 (69.2%)

The number of IP addresses and ASNs that the SIE DNS data observes DNS responses coming from over the course of two days.

The results the delay between domain registration date and date of first observed valid activity follow a long tail pattern. The majority of domains experience their first activity within two days of their registration. There is a tail in both directions from the registration time, with (%) of domains being subject to a valid DNS query before they were registered. The only exception to the smoothness of the long tail is an unexplained increase in the range 79-106 days after registration. This is centered around anomalous DNS activity that was observed on the days of January 13-14, 2011. On those days the number of domains that were observed to be successfully resolved for the first time was an order of magnitude higher than any other day in the range of SIE observation from August 1, 2010 to January 23, 2011 when collection ceased. The source of this anomaly remains unknown, but it does influence the distribution and average latency times for the baseline case of domain activity.

Of the 2,783,497 domains registered in October 2010, 2,064,091 (74.2%) were observed to have been referenced in the SIE DNS traffic in the observation window of August 1, 2010 to January 23, 2011. The majority of all observed domain names, 52.9%, is resolved within 1 day of the day they are registered. However, relatively few, 4.7%, are resolved on the same day they are registered. The number of domains that becomes active a given number of days after registration decays logarithmically.

The registration data for the domains the malware attempts to connect to can be partitioned in multiple ways. These different partitions can yield significantly different interpretations of the data. On the one hand, if one considers the domains related to code analyzed in Q4 2010, there were 146,856 unique domains. The SIE DNS database only had zone time data for 4,729 of these domains. On the other hand, one could consider the domains relevant to the artifact catalogue that happen to have been registered in October. This approach makes 504 domains available for analysis. In the

first case, an statistically indistinguishable percentage (33.2%) of the observed malicious domains is first resolved in the timeframes 50-95 days after they are registered and on the same day they are registered. In the case of October-registered malicious domains, a large majority, 73.0%, are observed on the same day they are registered.

Table II displays the delay time between registration of a domain and the first DNS response observed for all second-level domains for which we have registration data. It is expressed as both a percentage of all of the domains that were registered in October 2010 and as a percentage of those domains for which DNS messages were observed. Table III contains similar data for the subset of domains observed to be queried by malware in the malware database. Table III partitions this data both on those domains analyzed during Q4 2010 and the subset of those domains which were registered in October 2010. For both tables, the 99% confidence interval for the observed data is presented in parentheses.

TABLE II
DNS REQUEST DELAY FOR DOMAINS GENERALLY

Days	% of domains registered	% of domains observed
90-10 prior	1.8%	2.4% (2.4%,2.4%)
10-0 prior	1.4%	1.9% (1.9%,1.9%)
Same day	3.5%	4.7% (4.7%,4.7%)
1	35.8%	48.2% (48.1%,48.3%)
2	15.8%	21.3% (21.2%, 21.4%)
3 - 10	11.3%	15.2% (15.1%, 15.3%)
11 - 50	2.7%	3.7% (3.7%, 3,7%)
50-95	1.7%	2.2% (2.2%, 2.2%)
95+	0.3%	0.4% (0.4%, 0.4%)
Not observed	25.8%	N/A

Results for the general population of domain names in biz, com, info, mobi, and net registered October 1-31, 2010. Values in parentheses indicate the range for the .99 confidence interval of the observed data.

TABLE III
DNS REQUEST DELAY FOR MALICIOUS DOMAINS

Days	% of domains observed in Q4, all registration dates	% of domains observed with October registration dates
90-10 prior	0%	0.0%
10-0 prior	1.1% (0.7%,1.5%)	0.8% (0%,1.8%)
Same day	33.2% (31.4%,34.9%)	73.0% (68.0%,78.0%)
1	9.6% (8.5%,10.7%)	19.4% (15.0%,23.9%)
2	1.7% (1.2%,2.2%)	1.8% (0.3%,3.3%)
3 - 10	3.0% (2.4%,3.7%)	1.6% (0.2%,3.0%)
11 - 50	10.4% (9.2%,11.5%)	3.2% (1.2%,5.2%)
50-95	33.2% (31.5%,34.9%)	0.2% (0%, 0.7%)
95+	7.8% (6.9%,8.8%)	0.0%
Not observed	N/A	N/A

Results for the malicious population of domain names observed Q4 2010. Only reports on two-label domains for which there were zone time observed values in the SIE DNS database. Values in parentheses indicate the range for the .99 confidence interval of the observed data.

Table III does not relate data in regards to the total number of domains observed from the artifact catalogue. This would have cluttered the data over much because of the number of domains with features that the SIE DNS database does not have registration data for and factors associated with malicious domain behavior. Many pieces of malware look up a vast number of domains, only very few of which are intended to be resolved. Malware which exhibited this behavior is excluded from the sample of domains so that it would not overly bias the sample. It is excluded on the basis that if one piece of malware associated with a single MD5 hash is associated with 250 or more domain names, that MD5 and those domains names were not included in the study. This reduced the number of unique domain names in question from 146,856 to 33,795. Of these, 9,872 contained only two labels. Since both our zone file data set and the SIE database's zone file data contain only top level domains, the only domain names that could be in the data are those with two labels. Of these 9,872 two-label domains, 4,729 were found in the SIE database to have an entry for the time first observed in a zone file. The percentages for domains observed in Q4 in table III are calculated from these 4,729 domains.

The behavior between the set of domains obtained from malware and the set of domains generally is significantly different. The only case in which the 99% confidence intervals overlap is between the general case and the case of domains observed in Q4 which were registered in October. Between these two sets, the time interval of 11-50 days has overlapping ranges. 3.7% of domains generally fall in to this range, while between 1.2 and 5.2% of artifact-catalogue-related domains registered in October fall in to this range. All other latency bins between the general registration's latencies and the activity latency of the malware exhibited statistically significant differences.

IV. DISCUSSION

The domains represented by those from the artifact catalogue represent a particular type of malicious domain. In general, these are attempts to connect with a command and control server or drop box. These are essentially surreptitious activities. As such, they would be expected to behave differently than a phishing or drive-by-download malicious site, and this research should not be conflated as to seem to present a picture of all malicious domains. The malicious domains analyzed also represent a much smaller sample size, and we have no clear way to understand the bias that our sample may have. This is unlike the SIE data, which we have at least made an attempt to clarify the sampling bias.

Since the number of domains not observed varies so widely between the general and malware domains, it is more instructive to compare the time delays based on the percentage of the domains actually observed for a given delay. Malicious domains demonstrate a significantly different pattern of observed request delays than the general domains for every set of times. However, even within our one sample set, it is not clear in which way the malicious domains differ. In one case, they are activated much more quickly, and in another case they are activated much more slowly.

Further research will be necessary to better describe the domain life cycle patterns of different types of malicious domains. Once this research is accomplished, it could be utilized in order to help prevent the use of malicious domains before they become active. If many of the malicious domains are inactive for a long period of time after their registration, proactive registrars could keep track of which domains are

utilized after they are registered and potentially make de-registration decisions based on this data. The current data does not currently support in which manner that de-registration decision should be made. This research does indicate that a domain's pattern of activity in DNS traffic after its registration date is a valid area to search for such differences.

The comparison presented here is only an example comparison. Future work hopes to track how this distribution tends to change over time within the general internet domain population over time, among a wider set of TLDs, and with different sets of potentially malicious domains.

V. Conclusion

We believe that the information about the standard resolution patterns of domain names is potentially of utility to anyone performing analysis of DNS behavior. We hope that this baseline information can continue to be updated and standardized such that other researchers will be able to build upon this information. Furthermore, the indication that malicious domains resolve significantly differently than the average domain from their time of registration gives security researchers and domain managers another datum with which to attempt to identify and prevent malicious domains from causing damage.

References

[1] "Welcome to SIE." Internet Systems Consortium. <https://sie.isc.org/> 1/20/2011.

[2] J. M. Spring. "Large Scale DNS Traffic Analysis of Malicious Internet Activity with a Focus on Evaluating the Response Time of Blocking Phishing Sites," Master's Thesis, School of Information Science, Univ. of Pittsburgh, Pittsburgh, PA, 2010, pp. 26.

[3] C. Cohen, J. Havrilla, "Malware Clustering Based on Entry Points", CERT Research Annual Report 2008, pp 80. <www.cert.org/research/2008research-report.pdf>