An Internet-Wide View into DNS Lookup Patterns

Shuang Hao, Nick Feamster, Ramakant Pandrangi*
School of Computer Science, Georgia Tech *VeriSign Corporation

ABSTRACT

This paper analyzes the DNS lookup patterns from a large authoritative top-level domain server and characterizes how the lookup patterns for unscrupulous domains may differ from those for legitimate domains. We examine domains for phishing attacks and spam and malware related domains, and see how these lookup patterns vary in terms of both their temporal and spatial characteristics. We find that malicious domains tend to exhibit more variance in the networks that look up these domains, and we also find that these domains become popular considerably more quickly after their initial registration time. We also note that miscreant domains exhibit distinct clusters, in terms to the networks that look up these domains. The distinct spatial and temporal characteristics of these domains, and their tendency to exhibit similar lookup behavior, suggests that it may be possible to ultimately develop more effective blacklisting techniques based on these differing lookup patterns.

1. Introduction

The Domain Name System (DNS), the Internet's lookup service for mapping names to IP addresses, provides a critical service for Internet applications. The prevalence of DNS lookups can help network operators discover valuable information about the nature of the domains that are being looked up. In particular, the ability to perpetrate malicious activity on an Internet scale, ranging from spamming to scam hosting to botnet command-and-control depends on coordinating a large collection of hosts to perform a particular activity. Operators of these large-scale operations typically use the DNS to help direct hosts to the appropriate location on the network. In the case of various attacks such as scam and phishing attacks, these domains may be used to direct victims to a Web site (or through a proxy) that is hosting malicious content. In the case of botnet command-and-control, bots may locate the "controller" machine according to its domain name. Hence, the ability to identify domain names that correspond to malicious activity or otherwise unwanted traffic could be extremely valuable. The ability to characterize the behavior of these domains may not only help identify domains that are common to malicious behavior, but they may also help identify individual attacking or victim hosts.

Other studies have characterized DNS lookup behavior from different vantage points, such as below the local recursive resolver within an organization [1]. This previous study recognized that hosts within a single enterprise may exhibit coordinated lookup behavior to malicious domains, so clustering their activity patterns may yield information about the reputation of individual domains. Such a view of DNS lookup behavior is valuable, but this vantage point cannot reveal coordinated behavior across multiple networks. Because malicious activity often relies on the coordinated activity *across multiple networks*, the view of DNS lookup behavior below a single recursive DNS resolver may not completely capture behavior that is unique to malicious domains but is only visible from a perspective that captures lookup behavior across networks.

Towards the goal of understanding Internet-wide DNS lookup behavior for different types of domains, we study the spatial and temporal DNS lookup patterns for different domains across multiple networks. In particular, we characterize lookup patterns based on a cross-network view of these lookup patterns; this perspective allows us to see the IP subnets that host recursive resolvers that issue a lookup for a particular domain. This global perspective, combined with information about when a particular domain was registered, allows us to characterize both the spatial and temporal lookup patterns for individual domains (essentially, "Which network is looking up what, and when?") and identify characteristics that may be unique to domains that are somehow associated with malicious behavior. In this paper, we seek merely to *characterize* DNS lookup patterns for different types of domains and identify behavior that may be unique to domains that are associated with various types of attacks; it may be possible to develop sophisticated detection techniques (both for domains, and for individual IP addresses) based on the characteristics we identify, but this is not the focus of this paper.

To characterize DNS lookup patterns across networks, we use information about DNS lookups collected from the VeriSign top-level domain servers, coupled with registration information about these domains. These two pieces of information allow us to perform a study of DNS lookup patterns that is unique in two important ways. First, our analysis is based on a much more *global view* of DNS lookup patterns, as opposed to a view from any single network, which allows us to characterize the *spatial* properties of DNS lookups across the Internet. Second, we perform a *joint analysis* with information about when the domains were registered to explore the *temporal* properties of DNS lookups after their registration time.

We study spatial and temporal characteristics for both long-lived domains (Section 4) and newly registered domains (Section 5). We find that lookup patterns for scam, phishing, and botnet domains differ markedly from legitimate domains. Our study reveals the following findings:

- *Domains associated with scams and botnets exhibit more churn, in terms of the networks that look them up from day-to-day.* Although we see that the set of networks looking up any domain does vary from day-to-day, this variability is much more irregular for domains that are associated with spam and phishing attacks. This "churn" might serve as an important feature for quickly identifying malicious domains.
- *Domains that exhibit similar spatial lookup patterns also exhibit other similarities.* We cluster domains if they are looked up by similar groups of networks and explore whether these clusters contain domains of a similar type. We find that many such clusters contain either all legitimate or all "bad" domains, and that about 75% of all clusters contain domains that are at least 90% legitimate or 90% bad, which suggests that good and bad domains have distinct lookup patterns.
- *Blacklisted domains are typically queried by a much wider range of subnets, particularly for newly registered domains.* We explore the spatial diversity of networks that look up different types of domains and find that domains that appear in blacklists are queried by thousands of distinct /24 subnets, often within a few days of when the domain was registered. This varies markedly from the behavior of newly registered legitimate domains, which become "popular" less quickly.

We believe that these features can ultimately be used to develop unique network-wide lookup "footprints" that might also help distinguish legitimate domains from bad ones.

The rest of this paper is organized as follows. Section 2 surveys related work on using passive DNS monitoring to develop reputations for DNS domains. Section 3 describes the data sets that we use for our analysis. Section 4 explores the spatial and temporal characteristics of long-lived domains, whereas Section 5 describes these characteristics for newly registered domains.

2. Related Work

We survey related work in several areas: studies of DNS lookup behavior below the the recursive DNS server, other work on DNS reputation (including DNS-based blacklisting), and various studies of DNS lookup behavior at the DNS top-level domain servers.

DNS lookups patterns at local resolvers Perhaps the most related work to this study are the various studies of DNS lookup behavior below the resolver. The first studies of DNS lookup behavior at a local resolver were performed by Danzig *et al.* [5] and Jung *et al.* [7]; both of these studies examined lookup behavior from the vantage point of lookups to a single local resolver, and did not attempt to characterize how these lookup patterns differed by domain type. In particular, the Notos project [1] studies DNS lookup behavior within a local domain, below the DNS resolver, and attempts to distinguish good domains from bad domains based on the lookup patterns of clients that are issuing lookups to that resolver.

type	*example*
DNZA entry	add-new example.com NS ns1.example.com
Query record	example.com 111.111.111.0 , 22.22.22.0 3

Table 1: Data format examples.

Sato *et al.* apply similar techniques to analyze DNS lookup behavior below the recursive resolver and cluster hosts based on their DNS lookup patterns to identify bad domains and compromised hosts [10]. Both of these approaches aim to build DNS reputation by analyzing lookups as seen at a local recursive resolver. In contrast, this paper studies DNS lookup patterns from the perspective of a *top-level domain*, and examines the behavior of lookups as seen from recursive resolvers, as opposed to lookups from individual hosts.

DNS lookup patterns at root servers Other previous work has examined DNS lookup behavior as observed from a DNS root server [2–4]. The focus of these studies was much different from that in this paper. Castro *et al.* [4] and Brownlee *et al.* [3] attempt to characterize how much DNS traffic at the DNS root server is illegitimate, as a result of misconfiguration, typographical errors, etc. Broido *et al.* identifies misconfigured hosts using spectrography to identify hosts that are periodically (and mistakenly) issuing automatically configured DNS queries [2]. These studies look at query patterns at a root server, *not* at an authoritative top-level domain server. Additionally, they do not attempt to characterize how DNS lookup patterns vary according to their type.

DNS-based blacklists Various services attempt to build reputation for various DNS domains, either for helping users identify phishing domains (e.g., Phishtank [9]) or botnet command-and-control domains (e.g., Zeustracker [11]). In this paper, we do not attempt to directly infer the reputation of DNS domains. Some of the characteristics that we observe about various types of malicious domains may be useful for helping other dynamic reputation systems (e.g., Notos [1], Proactive Domain Blacklisting [6]) automatically identify whether a domain is being used for a particular activity.

3. Data

We provide a brief overview of our data set and then describe our process of labeling various domains in the dataset.

3.1 DNS Data

The top-level domain servers are responsible for maintaining the zone information (usually second-level domains) and answer the queries for the registered domains. VeriSign, Inc. operates the generic top-level domains (gTLDs) for `.com` and `.net`. The servers maintain two kinds of dynamics about the second-level domains. The first type of information is the *Domain Name Zone Alert (DNZA)*. This information includes changes about the zone, for example, whether a domain name was newly registered or a name server's IP address was modified. The DNZA files keep track of these changes these changes.

The second type of information concerns the *DNS queries* submitted by the recursive servers. Once the recursive servers sent queries to the TLD servers for resolving the 2LD domains names, the source IP addresses were aggregated into /24 subnets and the TLD servers monitored the number of queries for each domain. The query records show the relationship between the domain names and the queriers. The format examples are shown in Table 1. The DNZA entry showed that an "add-new" command created a new domain `example.com` and the NS record was `ns1.example.com`; The query record meant that there were 3 queries from /24s of "111.111.111.0" and "22.22.22.0" for the domain. The DNZA files and the query data were collected at VeriSign's `.com` and `.net` TLD servers during the period of October and November 2009. On average, there were about 80 million domains got queried each day.

To investigate the querying patterns for the domains, we identified several sources of "bad" domains and used them to label VeriSign's `.com` and `.net` domains. Due to the huge number of domains and the incompleteness of the labelling, the domains were also sampled for analysis. The domains were generally classified as two categories: "malicious domains" and "legitimate domains", and were further divided into the following subgroups.

3.2 Categorizing Domain Types

We now describe how we categorize domains according to different types, for both legitimate and malicious domains.

3.2.1 *Malicious domains*

Spamming We used a spam trap domain to capture emails sent from spammers during the period of July 2009 to January 2010. Since the domain was faked, no legitimate email messages would be destined to the domain. The emails received at the mail server were all spam. The second-level domains appearing in the messages' URLs were extracted as being involved in spamming activities. However, many popular legitimate domains, like `youtube.com`, were mixed in the data. To reduce the bias, we used the "popular legitimate domains" (in Section 3.2.2) as a whitelist to filter out the benign domains. Totally $29,363$ `.com` and `.net` second-level domains were considered spamming-related.

Phishing PhishTank [9] is a Web site that blacklists phishing sites. We crawled the published phishing URLs from February 2010 to April 2010. The domains showing in the URLs were blamed for hosting the phishing sites. It is noticed that although the data collection time was after November 2009, many of the URLs were verified to be phishing before October 2009. There were $6,230$ `.com` and .net second-level domains hosting phishing sites.

Malware or botnet-related domains Various DNS blacklists (e.g., ZeusTracker [11] and Malware Domain List [8]) were crawled during July 2009 to April 2010. The domains were detected to be connected by the infected machines, or to be used for fast-flux purpose. There were $10,952$ `.com` and `.net` second-level domains reported in the DNS blacklists.

3.2.2 *Legitimate domains*

It is difficult to identify legitimate domains, since people usually keep blacklist about malicious domains, but pay less attention to construct "whitelist" to mark legitimate domains. We used the following two methods to get representative sets of legitimate domains.

Popular legitimate domains Alexa (`www.alexa.com`) measures the web traffic volume and ranks the most visited domains. We collected the ranking data from February 2010 to April 2010. The top $10,000$ domains were changing over days. Malicious domains might cause sudden huge traffic and get high ranking, but we assume such bad domains will be detected very soon and get blocked. To get a "clean" list of popular legitimate domains, we identified the domains which always in top $10,000$ during the four months. $5,511$ domains are classified as popular legitimate ones.

Legitimate long-lived domains To get a representative set of legitimate domains, we sampled $10,000$ domains which get queried both at October 1, 2009 and November 30, 2009 (the first and last days over the course of our measurement study). Some restrictions are that the query number was greater than a threshold (set as 20), and the domain was neither in "popular legitimate domains" nor any malicious domain lists above. We assume the sampled domains were active during October and November 2009, had certain amount of visits (but not very popular) and were legitimate.

3.2.3 *Newly registered domains*

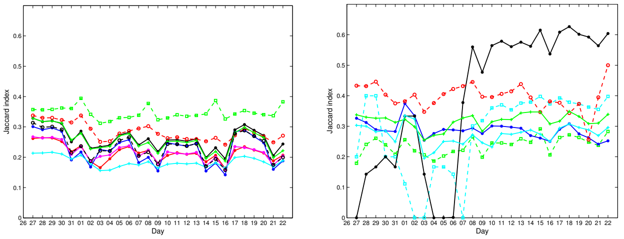
An orthogonal category is the set of newly registered domains (i.e., those that were registered within a short period of time before the timeframe of our dataset, or during our dataset). If different types of domains exhibit different query patterns shortly after registration, people might be able to spot the bad domains very quickly. Checking the "add-new" command in the DNZA files can identify $5,711,602$ newly registered domains during October and November 2009.

4. Characteristics of Long-Lived Domains

We explore the temporal and spatial characteristics of *long-lived domains* (i.e., those domains that have been registered for a relatively long period of time) and see how these characteristics differ across different types. We find that malicious domains see lookups from a much different set of networks from day-to-day (Section 4.1), and that domains that are used for similar purposes also exhibit spatial similarities (Section 4.2).

4.1 Temporal Behavior

On each day, the IP addresses that query a particular top-level domain can be represented as a set. We were interested



(a) Legitimate popular domains.

(b) Botnet C&C domains

Figure 1: Examples of the evolution of the lookup similarity index over time for various legitimate domains and botnet command-and-control domains.

to see how the set of domains that queried a particular domain evolved over time. We used the Jaccard index to measure the similarity of these sets over time. For two sets A and B, their Jaccard index is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

This metric measures the similarity between sets (the value range is among [0,1]). To compute a *lookup similarity index*—which we informally define as the similarity of the set of /24s that look up a domain from one day to the next—we aggregated IP addresses by their corresponding /24 network block and compared the /24 networks that looked up a particular domain for one day versus the next.

Figure 1 shows how this lookup similarity index varies over time for the set of /24 networks that query a domain on one day versus the previous day. We just show a very small subset of domains for demonstration. The x-axis represents the date (only the day, and the month is ignored). The time period is from October 26 to November 22, 2009. The y-axis is the Jaccard index, whose range is from 0 to 1. Figure 1(a) shows the Jaccard index for cases of *legitimate popular domains*, and Figure 1(b) shows the Jaccard index for cases of *botnet C&C servers* (a sub-category of the malware domains). Each line represents a domain, and a point on a line for a given day represents the lookup similarity for that domain for that day; due to privacy concerns, we do not label the domain names that correspond to each line.

The results show that this lookup similarity index for legitimate domains is considerably more consistent than the lookup similarity for the malicious domains. Some botnet command-and-control domains even sometimes have *no* overlapping /24 networks that query the domain from one day to the next. Upon further examination, we saw that these domains were domains that received very low volumes of queries on one day and a much higher volume of queries on the next. A likely explanation for this variance is that a botnet command-and-control domain may suddenly become popular and receive a high volume of queries on one day if a large fraction of hosts from new IP address regions suddenly become compromised.

To measure the variability of the Jaccard index for different categories of domains, we compute the coefficient of variance of the daily lookup similarity values over the period of October 2009 (31 days). Figure 2 shows the distribution of the coefficient of variance for similarity metrics for different types of domains. A larger coefficient indicates more "churn" from one day to the next, in general. The graph shows that the legitimate domains generally have significantly less churn in this similarity metric, while the malicious domains exhibit considerably more dynamics. This behavior could be useful for helping operators of authoritative domains identify bad domains of various types based on the churn that they exhibit.

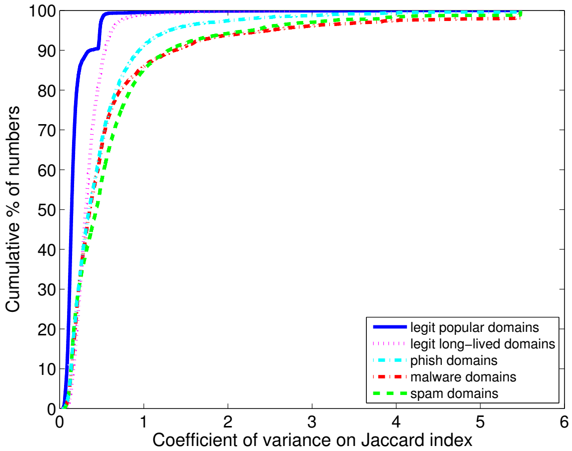


Figure 2: CDF of the coefficient of variance for lookup similarity. Legitimate domains exhibit considerably less day-to-day "churn" in the networks that look them up than malicious domains do.

4.2 Network-Wide Patterns

We next investigat the querying patterns *across different domains*, to see whether similar sets of networks were looking up different domains. Our intuition is that domains that are used for malicious purposes may be looked up by similar groups of networks as well. For example, a user cliking on a phishing URL might click other phishing URLs as well. If two domains share the same set of resursive DNS servers for querying, they are likely to belong to the same type of domains.

We measure the similarity using an average pairwise similarity of queried /24 network blocks over n days. Suppose two domains A and B who have sequences of quering /24 set $\{a_1, a_2, \ldots, a_n\}$ and $\{b_1, b_2, \ldots, b_n\}$ over n days. The similarity between domains A and B is caculated as

$$S(A, B) = \frac{\sum_{i=1}^{n} J(a_i, b_i)}{n}$$

where $J(a_i, b_i)$ is the jaccard index of set a_i and b_i. Based on this pairwise similarity, we aggregate the domains into different groups using single-linkage clustering, a simple and efficient clusteirng method [12]. We sampled 500 domains at random from each type of domain (for a total of $2,500$ domains) and considered a 6-day time preiod from

spam	*malware*	*phish*	*legit long-lived*	*legit popular*	% dominant type
41	36	28	11	386	79.40%
1	0	3	6	0	60.00%
0	0	0	0	8	100.00%
0	0	0	0	7	100.00%
0	1	1	4	0	66.67%

Table 2: Five largest clusters where a "good" domain type was dominant.

spam	*malware*	*phish*	*legit long-lived*	*legit popular*	% dominant type
70	46	63	29	1	85.65%
0	15	0	0	0	100.00%
0	7	0	0	0	100.00%
0	0	6	0	0	100.00%
0	1	4	0	0	100.00%

Table 3: Five largest clusters where a "bad" domain type was dominant.

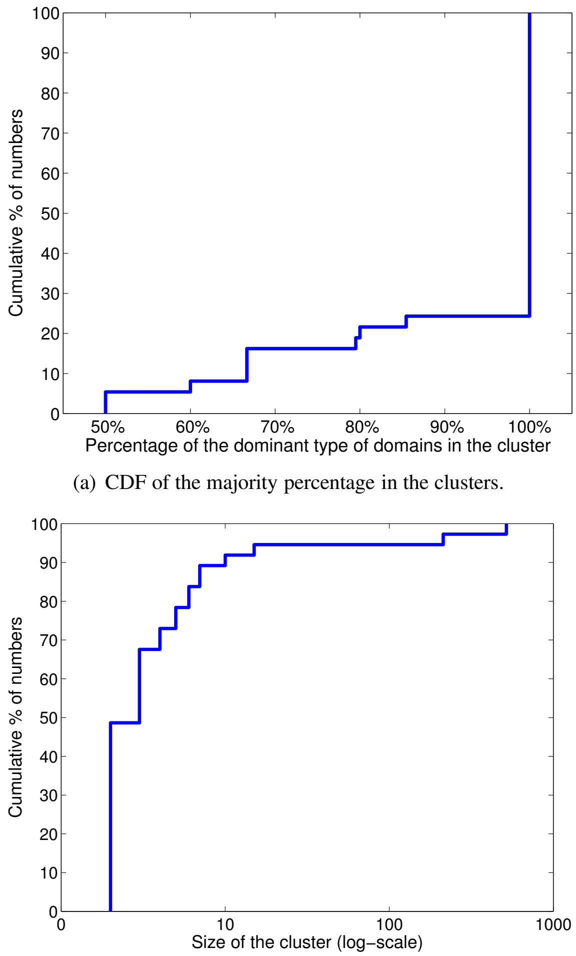
October 5–10, 2009. We terminate the clustering after $10,000$ comparisons, which places 700 domains into 39 clusters that have more than a single domain. We expect that domains of a similar type would fall into distinct clusters.

The cumulative distribution of the majority percentages of the clusters is shown in Figure 3(a); the y-axis represents the fraction of the cluster that is represented by the dominant domain type for that cluster. For example, if the cluster contained only phishing (or only legitimate popular) domains, then the domain would be 100% of the dominant type. Intuitively, the higher percentage that the majority type has, the "purer" the cluster is. This figure shows that around 75% of the clusters have a dominant type that represents more than 90% of the cluster, indicating that similar domain types do, in fact, exhibit similar network-wide lookup patterns across networks. The cumulative distribution of the cluster sizes is shown in Figure 3(b). Most of the clusters have small number of domains, less than 10.

Tables 2 and 3 show the statistics for the five largest good and bad clusters. The columns represent the numbers of "spam", "malware", "phish", "legitimate long-lived" and "legtimate popular" domains in the clusters. The last column shows the percentage of the cluster that is represented by domains that are either "good" or "bad". We define "bad" domains as the aggregation of column 1 to 3; and "good" domains are the sum of column 4 and 5. The percentage shown is the portion of the cluster that is represented by domains of the dominant type. The results show that, in many cases, the clustering works quite well: many of the clusters contain either only all good or all bad domains. These preliminary results suggest that domains of certain types do share similar network-wide spatial lookup patterns that may ultimately be used as input to a blacklist.

5. Evolution of Newly Registered Domains

This section characterizes lookup patterns for newly registered domains during October and November 2009. The "add-new"command in the zone DNZA files indicates the registration of domains; the records contained $5,711,602$ such newly registered domains during this period.



(a) CDF of the majority percentage in the clusters.

(b) CDF of the cluster sizes.

Figure 3: CDFs about the single-linkage clustering.

Our intuition is that once being deployed, malicious domains may receive a huge amount of traffic in a short time, but the visits to legitimate domains will climb slowly after registration. The period from October 1–26, 2009 had $1,647,964$ domains newly registered in total. Of these newly registered domains, 384 `.com` and `.net` domains were related with "malware" in blacklists, and $1,690$ domains appeared in "spamtrap" messages. Because the overlap with "phish" domains was very small, we do not perform any analysis for these "phish" domains in this section.

Figure 4 shows the a CDF of the number queries on October 26. The x-axis is number of unique /24s, and the y-axis shows the percentage of domains whose metric value less than or equal to the value on x-axis. The top curve (solid line) shows the distribution for domains not reported as malicious; the other curve shows the distribution for domains in "spamtrap" and "malware" types, respectively. These curves indicate that the bad domains attracted significantly more /24

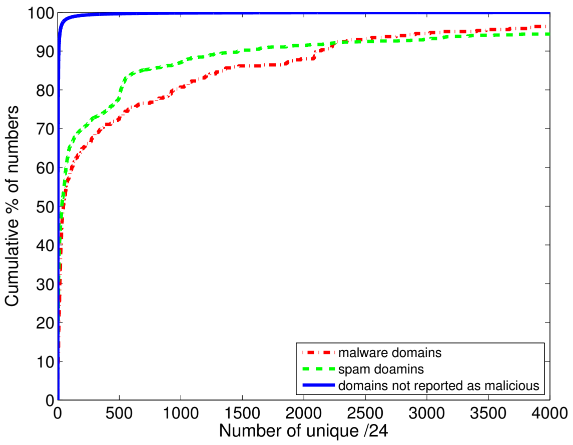


Figure 4: Distribution of /24s on October 26, 2009 for newly registered domains.

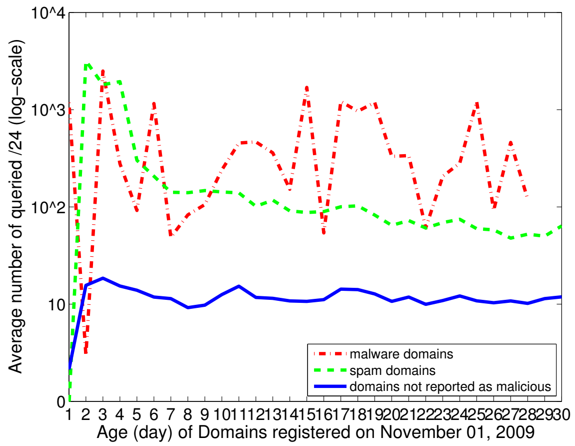


Figure 5: Evolution of querying number over time.

subnets within 25 days after registration. When checking the domains not reported as malicious, 80% of the domains are queried by fewer than 5 unique /24s, but the distribution curve has a long tail (some domains had as many as 10^5 querying /24s).

To study how the dynamics of queries evolved after registration, we focused on the domains registered on November 1, 2009, and inspected the querying /24s for the domains in the next 30 days (one month). There were 28 "spam" domains, 9 "malware" domains and 53, 105 unreported domains registered on that day. Figure 5 shows the average numbers of /24s for the domains in different categories over time. The x-axis shows the number of days after November 1, 2009 (the date that the domains were registered). The y-axis shows the average number of querying /24s over the domains in the same category (on a logarithmic scale). The queries to the bad domains increased quickly after the domains were registered. On the other hand, the /24s to query for the domains not reported as malicious increased slowly and stayed relatively low over the 30-day period. These markedly different lookup patterns of the likely legitimate domains and those involved in malicious activities might ultimately help blacklist operators quickly detect bad domains, by watching for newly registered domains that suddenly become popular.

6. Conclusion

We have analyzed DNS lookup patterns from a large authoritative top-level domain server and characterized how lookup patterns for "bad" domains (e.g., those used for spamming, phishing, malware and botnet command-and-control) may differ from those that are used to host legitimate domains. We examine domains for phishing attacks and botnet command-and-control domains and see how these lookup patterns vary in terms of both their temporal and spatial characteristics. We find that botnet and phishing domains tend to exhibit more variance in the networks that look up these domains. We also note that miscreant domains exhibit distinct clusters, in terms of the networks that look up these domains. Finally, we find that these domains become widely popular considerably more quickly after their initial registration time, both in terms of the number of distinct network blocks looking them up and in terms of total query volume. The distinct spatial and temporal characteristics of different types domains, and their tendency to exhibit similar lookup behavior, suggests that it may be possible to "fingerprint" domains based on their lookup patterns and ultimately develop more effective blacklisting techniques based on these DNS lookup patterns.

REFERENCES

[1] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster. Building a Dynamic Reputation System for DNS. In *Proc. 19th USENIX Security Symposium*, Washington, DC, Aug. 2010.

[2] A. Broido, E. Nemeth, and K. Claffy. Spectroscopy of DNS update traffic. *ACM SIGMETRICS Performance Evaluation Review*, 31(1):321, 2003.

[3] N. Brownlee, K. Claffy, and E. Nemeth. DNS measurements at a root server. *GLOBECOM*, 3:1672–1676, 2001.

[4] S. Castro, D. Wessels, M. Fomenkov, and K. Claffy. A Day at the Root of the Internet. *ACM SIGCOMM Computer Communication Review*, 38(5):41–46, 2008.

[5] P. Danzig, K. Obraczka, and A. Kumar. An analysis of wide-area name server traffic: A study of the internet domain name system. *ACM SIGCOMM Computer Communication Review*, 22(4):292, 1992.

[6] M. Felegyhazi, C. Kreibich, and V. Paxson. On the potential of proactive domain blacklisting. In *Third USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET '10)*, 2010.

[7] J. Jung, E. Sit, H. Balakrishnan, and R. Morris. DNS Performance and the Effectiveness of Caching. In *Proc. ACM SIGCOMM Internet Measurement Workshop*, San Fransisco, CA, Nov. 2001.

[8] Malware domain list. `http://www.malwaredomainlist.com/`.

[9] Phishtank. `http://www.phishtank.com/`.

[10] K. Sato, K. Ishibashi, T. Toyono, and N. Miyake. Extending Black Domain Name List by Using Co-occurrence Relation between DNS Queries. In *3rd USENIX Workshop on Large-Scale Exploits and Emergent Threats*, May 2010.

[11] Zeustracker. `http://zeustracker.abuse.ch/`.

[12] J. Zupan. *Clustering of Large Data Sets*, 1982.