The Effectiveness of Block Lists to Prevent Collisions

Matthew Thomas
Verisign Inc.
12061 Bluemont Way
Reston VA 20190
mthomas@verisign.com

Yannis Labrou
Verisign Inc.
12061 Bluemont Way
Reston VA 20190
ylabrou@verisign.com

Andrew Simpson
Verisign Inc.
12061 Bluemont Way
Reston VA 20190
asimpson@verisign.com

ABSTRACT

The Alternate Path to Delegation proposed by ICANN places a premium on the value and relative sample strength of NXDomain (NXD) requests within the "Day in the Life of the Internet" (DITL) datasets. The domains observed lead to the creation of an authoritative list of Second-Level Domains (SLDs) to be blocked from registration for all eligible gTLD applicants. However, the efficacy of this block list is still unclear. In this paper, we will examine the NXD request patterns observed within the DITL data and that of a longer longitudinal study of the A and J roots in an effort to better gauge the effectiveness of such a collision blocking technique.

1. INTRODUCTION

Constructing a block list or a blacklist to mitigate potential naming collision requires observational data in which an empirical judgment of risk mitigation can be calculated. Ensuring the underlying data sample is representative of the real-world risks is pertinent to guarantee the efficacy of such a blacklist. Limited data sources and/ or biased data may prove to be sub-optimal and result in ultimately a flawed mitigation strategy. Furthermore, it is often common to revise and manage such a list as time progresses and the monitored system's behavior evolves. The proposed Alternate Path to Delegation (APD) [1] relies on the accuracy and sample strength of data within the "Day in the Life of the Internet" (DITL) [2] datasets. The DITL data consists of DNS query and responses over a continuous 48-hour sample from various Root DNS operators for the past several years. APD requires a registry operator to initially block all second-level-domains (SLDs) that appear in the DITL data after applying an ICANN described methodology per top-level domain (TLD). Such a measure would prevent these SLDs from resolving and prevent potential name collision events. We will use the nomenclature of an SLD to be the combination of a unique label tied to a specific top-level domain TLD throughout the remainder of this paper (e.g. "example.tld").

The effectiveness of such a list is dependent on accuracy and completeness and sampling accuracy of SLDs observed within the DITL sample data. Such measurements are difficult to gauge due to the limited longitudinal data retained within the DITL datasets and the variable contribution of different root operators from year to year. In this paper, we will attempt to measure the technique of DITL-based block listing by first analyzing the historical data within DITL over the past seven years and second by using a longer longitudinal study of the A and J Root DNS servers. Insights drawn from the DITL data will potentially allow us to draw parallels and compare overall trends of DNS data and correlate those events with what is seen in Verisign operated A and J root servers [3] over a continuous five month period of time.

2. DATA PROCESSING

The following section describes the data used within the paper and all pre-processing and filtering techniques that were applied to the data.

DITL: The DNS-OARC Collisions group, including members Roy Hooper of Demand Media and Kevin White of JAS Global Advisors, did a large amount of mining the raw DITL data into smaller and more manageable files based on the new proposed TLD strings. These files were separated by TLD and by year allowing for easy analysis. The details of the files and their format can be found on the DNS-OARC site [4].

A & J Root: Verisign is the Root operator for both A and J. NXDomain (NXD) responses were captured from both root nodes starting July 16, 2013. This data was similarly transformed to a format analogous to that of DITL's transformation undertaken by Demand Media and JAS in which requests were stored by TLD and date and limited to the set of applied for gTLDs [5]. The data set consisted of approximately 3.6 billion NXD records and approximately 27.5 million unique SLD's.

TLD String Exclusions: As part of ICANN's risk mitigation strategy, the TLD's "home" and "corp" were deemed to be high risk due to substantial usage of those names internally by various entities and the abundance of NXDomain data at the roots [6]. Accordingly, we have excluded all requests of these TLDs in our figures and analysis.

SLD String Exclusions: Many DNS queries contained requests for randomly generated 10 character alphabetic strings, most likely used by Google Chrome to detect specific aspects of DNS resolver behavior [7]. In an effort to remove these queries, both the DITL and A&J data sets were filtered using a technique analogous to ICANN's APD filtering methodology described per TLD [8].

3. DITL OBSERVATIONS

Longitudinal SLD Growth

Using DITL data since 2006, the number of unique SLDs was measured for each year (Figure 1: Observed SLDs), as well as the number of SLDs seen in previous years, but measured against the current year's distinct set of SLDs (Figure 1: Previously Observed SLDs). The upward trend clearly shows a steady growth rate of new SLDs and that the increasing difference between Observed and Previously Observed measures has steadily increased over time. These initial observations may already indicate that the yearly delta increase between previously observed SLDs and new SLDs presents a potential problem with DNS-based block listing, as any list based on a limited sample of data in a highly evolving system will be misrepresentative.

Coordinating all DNS Root operators to contribute and participate in the DITL exercise is a difficult and sometimes impossible task. Many of the historical data points within DITL have limited or missing portions of data from specific roots. However, more recently the amount of participation has increased and could potentially contribute to the increasing observations noted in Figure 1.

Ideally, we would like to measure the growth rate between Observed and Previously Observed SLDs on a more frequent basis rather than DITL's 48-hour yearly snapshot. Given the difficulties of running a multi-root coordinated 48-hour capture, it is most likely implausible to conduct a longer longitudinal study with all of the roots. While conducting a longer longitudinal study at a subset of roots may prove to be feasible, it first requires that we ensure that the general trends and statistical significance of any one root, or subset of roots, is representative of the whole.

In order to measure the affinity any particular SLD has with a given root, the number of distinct roots a SLD has queried for a given DITL year is measured in Figure 2 (every year there are roots that did not contribute to DITL). The vast majority of SLDs interact with only one root, suggesting observational sampling at a specific root would be biased and of limited value for blacklisting purposes; however, this affinity may prove useful to study a SLD's longitudinal patterns by sampling from a specific root.

Figure 2 does not take into account the volume (number of queries) of each SLD. It is very likely that many of the SLD's are only queried a few times (or, once) over the 2-day window of the DITL data, thus making it unlikely that these few queries will hit multiple roots. Furthermore, evidence from Google's recursive DNS service suggests that the frequency and diversity of SLDs observed within DITL may be severely under represented due to recursive level NXD caching [11]. Figure 2 suggests that a limited 2-day sampling window from a subset of the roots will be a poor candidate for any blocking approach due to strong SLD-root affinities.

Figure 1 – Second Level Domains Observed in DITL by Year for Proposed gTLD Strings:

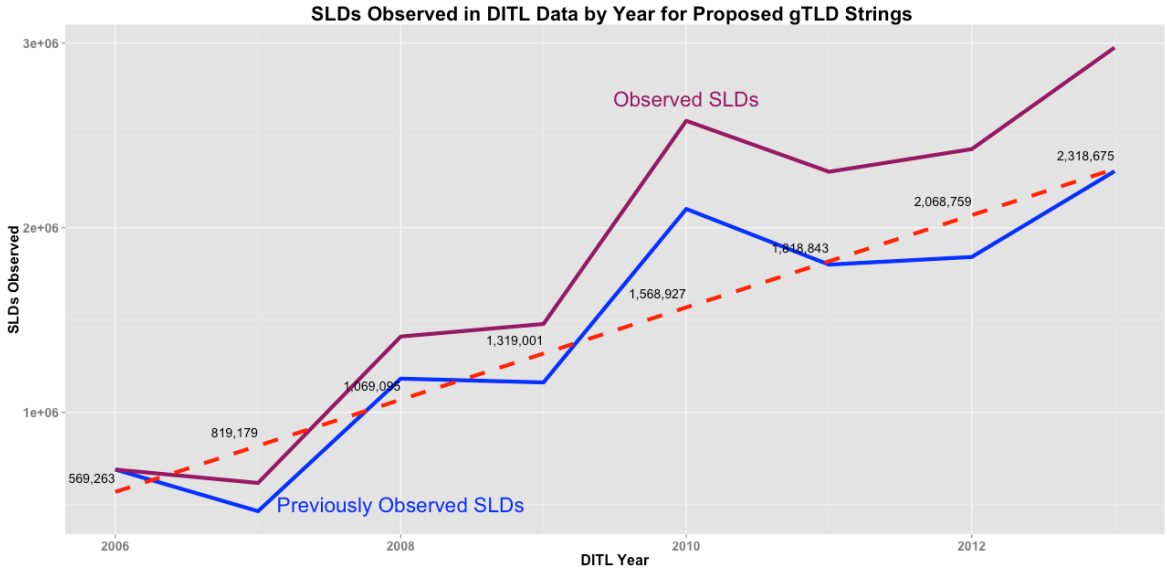
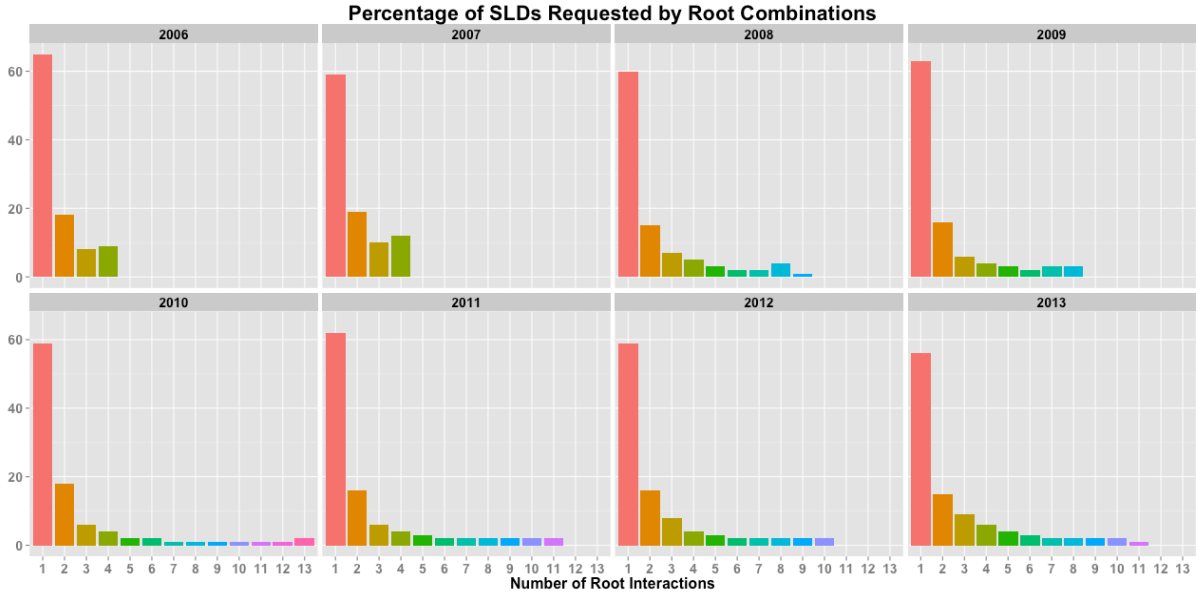


Figure 2: Percentage of SLDs Requested by Root Combinations



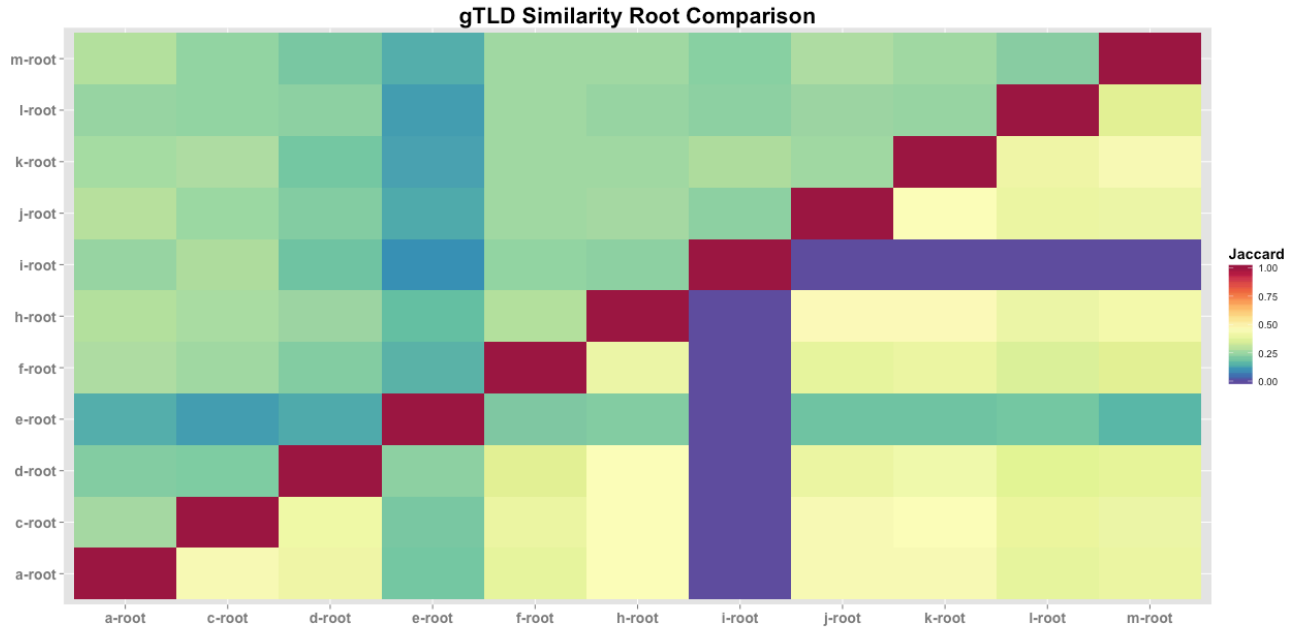
Root Affinity Measurements

In order to understand non-single root interacting SLDs, measurement affinities between each of the roots was conducted. Specifically, for each root pair, a similarity measurement using the Jaccard Index [9] was calculated. The Jaccard Index similarity is a very simple measure that reflects the intersection of two sets over their union. The resulting metric ranges from zero to one or no-overlap to identical sets.

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

Figure 3 focuses on two root affinity similarity measurements: 1.) Distinct SLDs observed (upper left half of Figure 3). 2.) Distinct /24 network requests (lower right half of Figure 3). I-Root traffic in the lower right half appears to have zero similarity due to the fact that I-Root anonymized source IP addresses because of privacy regulations with the country of that data's origin [10]. E-Root's abnormal behavior in both halves of the matrix is a bi-product of E-Root's smaller DITL data contribution and the Jaccard Index metric itself [10]. It is worth noting that both B and G roots are not present as the figure is based on 2013 DITL data set. Despite these outliers, there appears to be no inter-root affinity for either specific SLDs or recursive name server traffic.

Figure 3: gTLD Similarity of SLDs and /24 Networks at Various Roots



4. LONGITUDINAL INSPECTION USING A+J ROOT

Verisign is the Root operator for both A and J nodes. NXDomain responses were captured from both root nodes starting July 16, 2013 until Dec. 31, 2013 and the resulting 3.6B records were transformed to a format analogous to that of DITL's transformation undertaken by Demand Media and JAS.

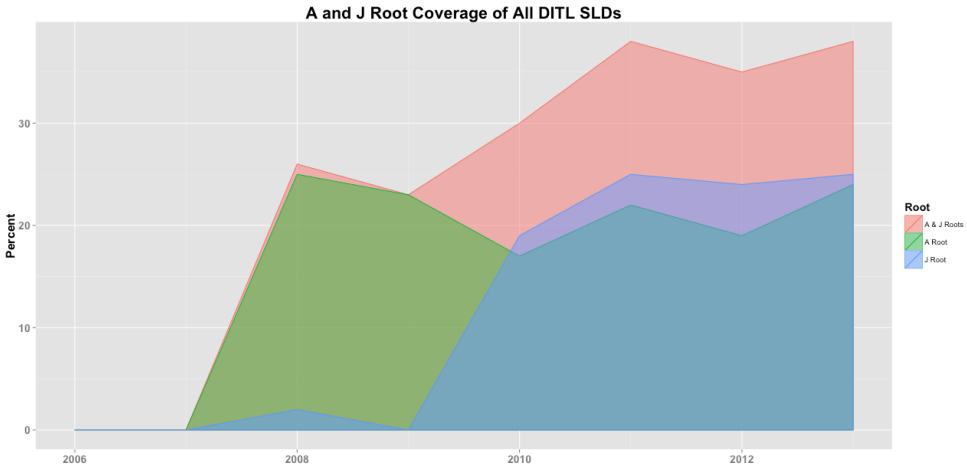
In the absence of sufficient longitudinal data in the DITL data set it is hard to access the effectiveness of the current blocking based solely on the DITL data. Hence, we intend to use the A+J Root NXD data as a proxy for the evaluation of the blocking strategy. The central questions we seek to answer are:

- How representative of the NXD traffic is a 2-day data sample from all the participating roots?
- How effective can a blocking strategy based on the DITL sample be?

We will next present our analysis of the A+J data with a focus on answering these two questions.

First, we will examine how representative the A, J roots are of the root NXD traffic overall. Based solely on the DITL data over the years, we can see (Figure 4) that the combined A+J roots, on an annual basis, observe a bit less than 40% of all the SLD's observed across all roots; each of A, J separately observe roughly 23%. Since the ratio is based on annual aggregates we postulate that A+J observe a significant percentage of all SLD's given a sufficiently long period (a year). The data also corroborates the data illustrated in Figure 3.

Figure 4: A and J Root Coverage of All DITL SLDs



We next turn our attention to the Verisign initiated constellation-wide collection of A and J Root NXDomain data beginning July 16, 2013. Figure 5 illustrates a longitudinal plot, over the collection period, of the number of unique SLDs observed on a given day and the number of previously seen SLDs prior to that day. The methodology is the same as that of Figure 1, which is based solely on the DITL data.

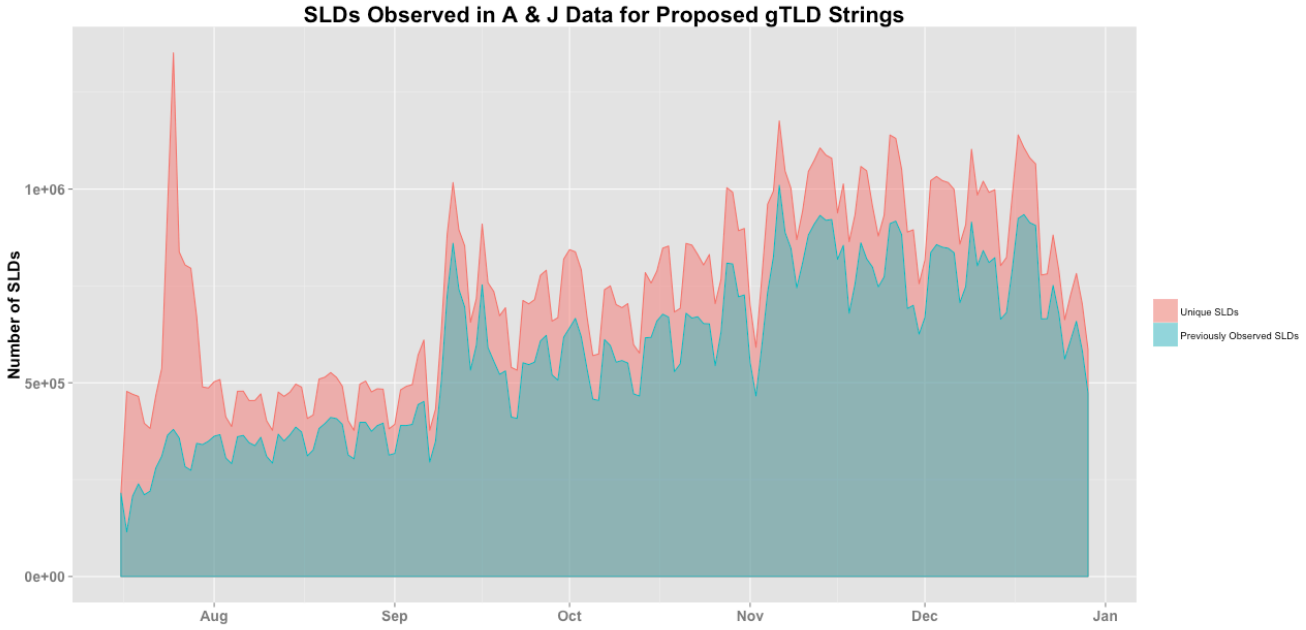
The trend clearly shows significant differences between these two measures, reaffirming that the daily change of SLDs within the applied for TLDs, even after Chrome filtering, is extremely high. The average percentage of newly seen SLDs for any given day over this period of time is 22.5%, or, on average, 22.5% of the SLD's observed by A+J roots on every single day have not been observed before, or at least not since the beginning of the collection period, by A+J roots. Figure 5 suggests an extremely high entropy in the universe of NXD SLD's and that any small collection window (1 or 2 days) will only account for a small percentage of the SLDs over the day, week or month following the collection. Perhaps, more alarmingly, the pattern is so consistent that any collection period (no matter how long) will always have a large number of never seen before SLDs in the subsequent days or weeks. Moreover, any method that relies on a small observation window cannot distinguish between those SLDs that have appeared before or appear for the first time.

There are two more interesting observations to add:

- The data exhibits the familiar weekly pattern (lower numbers on weekends) of DNS resolution activity

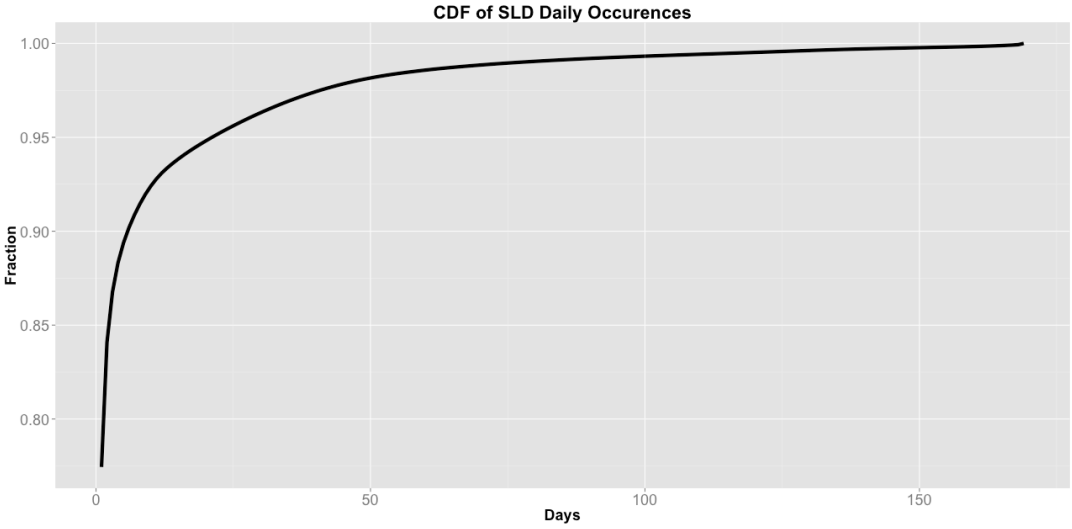
- It is possible that at least some of the SLDs are being generated due to awareness of the TLDs and blocking policies, thus artificially exaggerating the delta. We are not aware of any ways to gauge if said awareness influences the observed data.

Figure 5: SLDs Observed in A & J Data for Proposed gTLD Strings



A very significant percentage of these SLDs appear only briefly (and never seen again). Indeed, when inspecting the number of days during which these SLDs appeared during the measurement time period it turns out that the vast majority of SLDs appeared very few days. Figure 6 illustrates a Cumulative Distribution Function (CDF) of the number of days each SLD appeared during the observation period. For example, nearly 80% of the observed SLDs only appeared during one day in the 167-day measurement and only 5% of SLDs appeared on more than 20 days, or during 12% of the measured days. Still, 5% is a significant number of SLDs (~ 1.375 million) that have a consistent presence in the data but the finding should partially explain the high entropy observed in Figure 5.

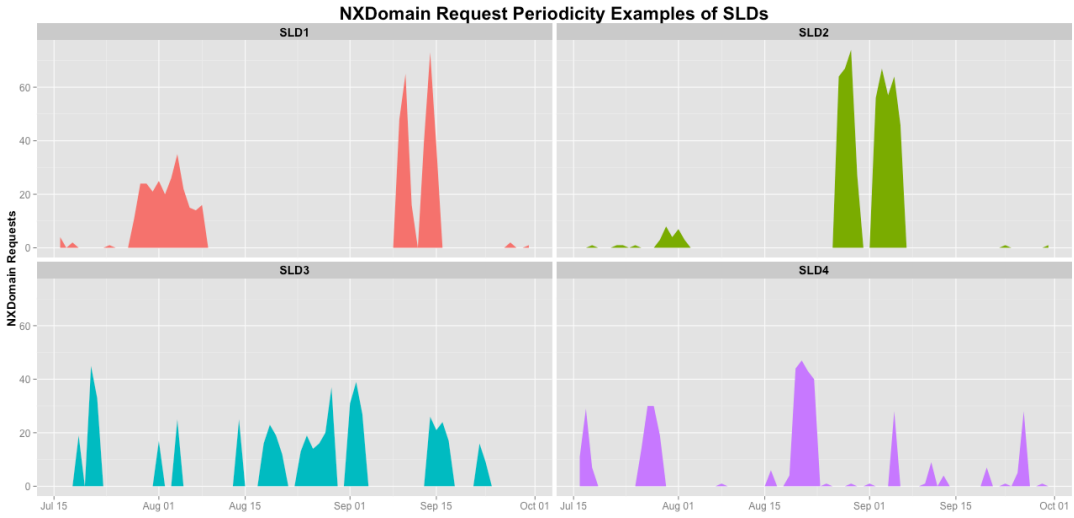
Figure 6: Cumulative Distribution Function of SLD Daily Occurrence Frequencies



Understanding the query periodicity for non-singular SLD observations is an important facet to consider, as it will help ensure the collection window used to create block lists is appropriately sized. Since a significant portion of SLDs occurs only on one day, understanding the inter-query time distribution of those multi day SLDs may prove to be insightful.

Figure 7 depicts the query patterns of 4 SLDs observed at the A&J roots. It is clear that these domains exhibit some form of "burstiness" in which many NXD queries are issued for a SLD and subsequent NXD queries are not observed for a significant period of time. If block listing data collection was performed during one of these SLD "quiet periods", the block listing technique would obviously under represent and miss potential SLDs with name collision potential. It is worth noting that macro Internet networking phenomena may account for this burstiness, but without longitudinal full root participation over a large time period it is impossible to speculate on the causes.

Figure 7: NXDomain Request Periodicity Examples of SLDs

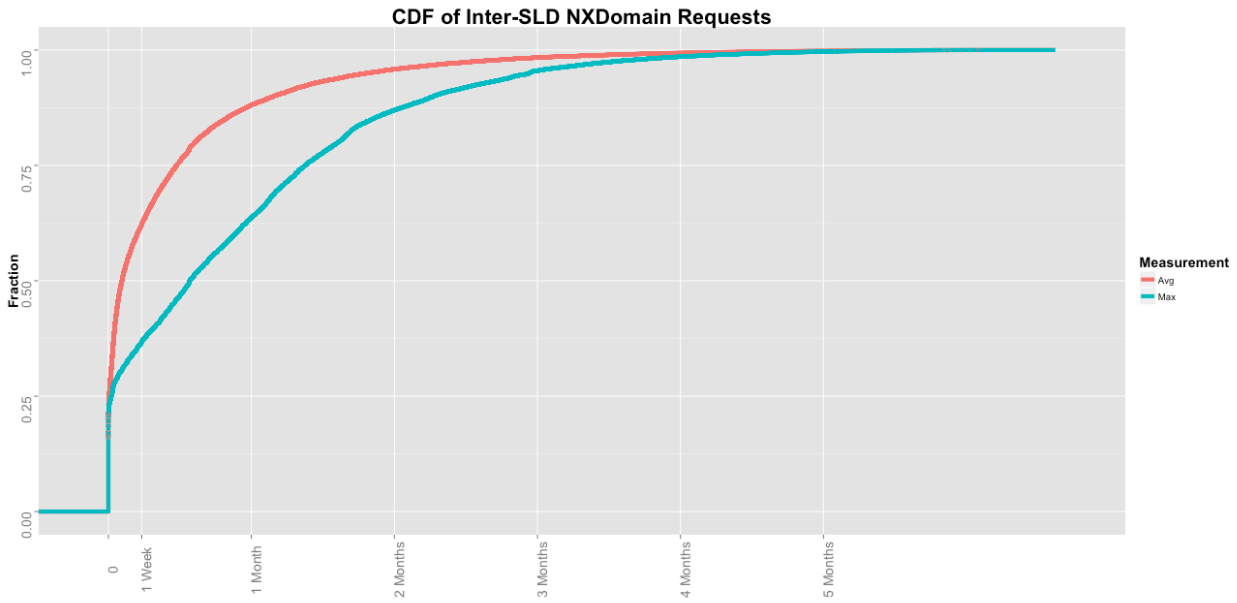


The previous figure suggests the presence of "burstiness", where the quiet time in-between queries is typically very large, but we would also like to quantify this behavior over the complete dataset. Figure 8 is a CDF plot of the average inter-query time distribution and of the maximum inter-query time distribution for all observed SLDs. These measurements were conducted as follows: for each SLD with at least 2 queries during the observation period, we ordered all the timestamps of the requests and measured the time between consecutive queries, and for each series of inter-query time intervals we calculated the average and the maximum value. For example, according to Figure 8, 37% of the domains (more than 10 million SLDs) the average inter-query period is one week and the largest (maximum) period of silence is 2.5 weeks, or, as another example, 12% of the SLD's (more than 3 million) might have no queries for as long as two months. Overall, Figure 8 suggests a very pronounced burstiness behavior.

Based on the empirical evidence of Figure 8, we want to next investigate the effect of increasing the collection window as a logical next step to address the burstiness we observed. The intent is that a larger window can capture a more representative set of SLDs for blacklisting purposes. Figure 9 is an effort to evaluate alternative collection window sizes and their potential effectiveness for "blocking" SLDs. We first calculate the set of blocked SLDs that are present in the data at the beginning of the observation period. Then, for each moving window of a size of N days we calculate the percentage of SLDs in the window that are also present in the initial blocked SLD's list, for every instance of the moving window over the observation period. Referring to Figure 9, each line represents a moving window size of N days

(from 2 to 14 days); the y-axis shows the percentage of SLDs that were in the blocked SLD's list for every date in the observation period (x-axis).

Figure 8: Cumulative Distribution of Inter-SLD NXDomain Requests

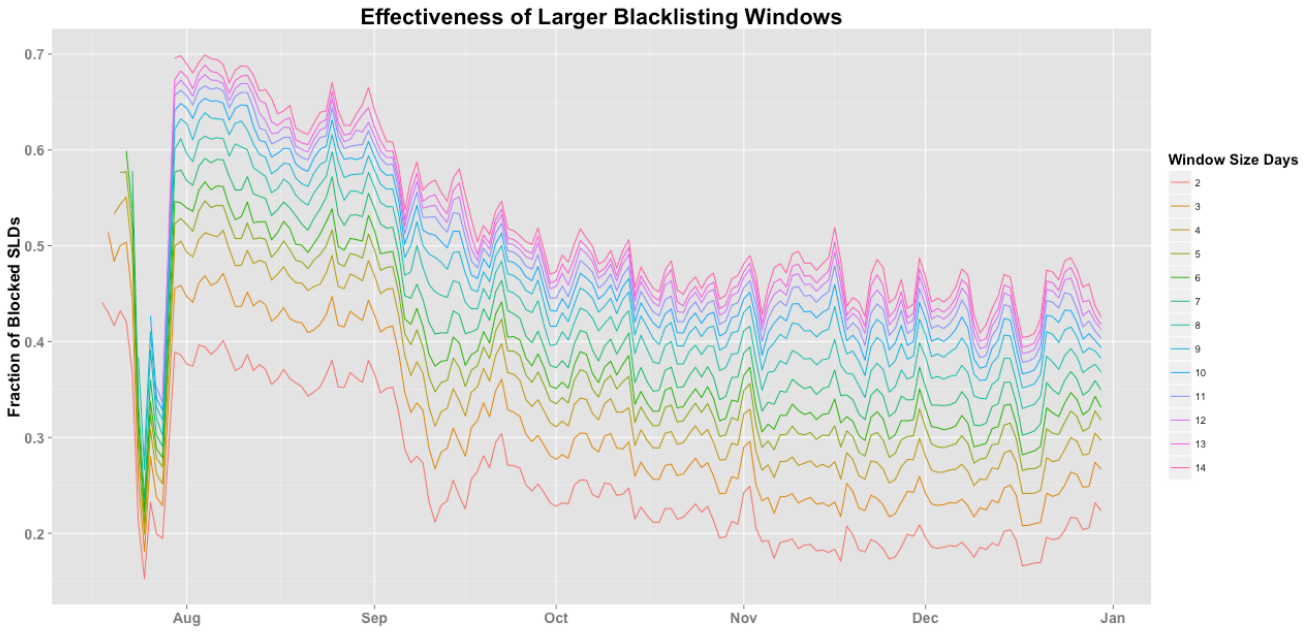


We can make the following observations:

- As the window size increases, the percentage of observed blocked SLDs also increases, but the effect of that increase becomes smaller and smaller as the window increases. This suggests that there are limits to the benefits of an increasing observation window, which seems to asymptotically approach an upper bound.
- For any given size window, the ratio becomes smaller as time progresses. This is hardly surprising given the significant number of never seen before SLDs that appear on every single day, as illustrated in Figure 5. A related conclusion is that as time progresses, the block list will become (perhaps asymptotically) less effective as it represents an increasingly smaller fraction of the observed SLDs. Of course, there is no way to tell if the any of the newly SLDs should have been blocked but since there is nothing unique about the days of the collection of the DITL data (which lead to the block list) there is no reason to believe that a significant percentage of the newly SLD's would not have been blocked.

Summarizing the investigation of the A+J root NXD data, we believe that the significant number of new NXD SLDs that appear on every single day, coupled with the very bursty behavior of requests for the same SLD, necessitate a larger collection window, but there are clear limits to the effectiveness of a larger window. One limitation of those results is that they are based on combining the observations of only 2 roots instead of processing longitudinal data from all roots (which are of course not available). We would argue that given the limited overlapping between roots (see Figure 3) it is reasonable to assert that our results would not be radically different if applied to longitudinal data from all roots.

Figure 9: Effectiveness of Larger Blacklisting Windows



5. CONCLUSION

We have investigated the potential efficacy of block listing domains in order to prevent name collisions by measuring various forms of NXDomain traffic recorded within the DITL dataset and that of a longer continuous period from the A and J root servers. Our findings suggest that even with the limited longitudinal data points contained within DITL, a highly dynamic and evolving SLD universe exists in which block listing based on static DNS samples will both over estimate and under estimate potential name collisions. The A and J root data constitutes a significant portion of the SLDs observed at the roots, and a longitudinal study of the A+J NXDomain traffic illustrates temporal patterns of bursts and churn that is not observable within DITL – further suggesting longer collection windows are required to ensure a representative sample is captured. Finally, the effectiveness of blacklisting domains by modifying collection timeframes or data sets will prove to be ineffective due to the highly entropic SLD universe, and an alternative methodology of mitigating collisions should be used in conjunction or in place of the current approach. For example, repeating the collection and evaluation of new and previously observed SLDs at regular intervals from all root nodes.

ACKNOWLEDGEMENTS

We thank DNS-OARC, ICANN, JAS and Demand Media for providing data and resources to help guide and conduct this study on name collisions.

REFERENCES

[1]

"Reports for Alternate Path to Delegation Published." ICANN New GTLDs. N.p., n.d. Web. 05 Feb. 2014.

http://newgtlds.icann.org/en/announcements-and-media/announcement-2-17nov13-en

[2]

"DITL Traces and Analysis | DNS-OARC." DITL Traces and Analysis | DNS-OARC. N.p., n.d. Web. 05 Feb. 2014.

https://www.dns-oarc.net/oarc/data/ditl

[3]

"Domain Name System." And What Is DNS. N.p., n.d. Web. 05 Feb. 2014.

http://www.verisigninc.com/en_US/domain-names/online/domain-name-system/index.xhtml

[4]

"2013 Collisions Project DITL Analysis | DNS-OARC." 2013 Collisions Project DITL Analysis | DNS-OARC. N.p., n.d. Web. 05 Feb. 2014.

https://www.dns-oarc.net/node/332

[5]

"ICANN New GTLDs." ICANN New GTLDs. N.p., n.d. Web. 05 Feb. 2014.

http://newgtlds.icann.org/en/

[6]

"Internet Corporation for Assigned Names and Numbers." Addressing the Consequences of Name Collisions. N.p., n.d. Web. 05 Feb. 2014.

http://www.icann.org/en/news/announcements/announcement-3-05aug13-en.htm

[7]

"ISC Diary." Isc Home. N.p., n.d. Web. 05 Feb. 2014.

https://isc.sans.edu/diary/Google+Chrome+and+(weird)+DNS+requests/10312

[8]

"Internet Corporation for Assigned Names and Numbers." Alternate Path to Delegation Report for .luxury. N.p., n.d. Web. 05 Feb. 2014.

http://www.icann.org/en/about/agreements/registries/luxury/luxury-apd-report-12nov13-en.htm

[9]

"Jaccard Index." Wikipedia. Wikimedia Foundation, 15 Jan. 2014. Web. 05 Feb. 2014.

http://en.wikipedia.org/wiki/Jaccard_index

[10]

"2013 DITL Data | DNS-OARC." 2013 DITL Data | DNS-OARC. N.p., n.d. Web. 05 Feb. 2014.

https://www.dns-oarc.net/node/325

[11]

"Re: Public Comments on Proposal to Mitigate Name Collision Risks, by Google Inc." ICANN Forums. N.p., n.d. Web. 05 Feb. 2014.

http://forum.icann.org/lists/comments-name-collision-05aug13/pdfkwCAlijJOp.pdf