Analysis Techniques for Determining Cause and Ownership of DNS Queries

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

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

♦Equally Contributing Authors

ABSTRACT

Authoritative name servers observe a large amount of DNS queries for namespaces that are not delegated. While it may not be clear why a resolver receives these requests, the possibility of naming collisions may occur if the names were to be delegated. Therefore, understanding the context and intention of such queries may prove useful to better assess the amount of risk associated with such delegations. Unfortunately, DNS queries contain a relatively small amount of information, as the protocol was designed for efficiency and limited bandwidth usage – making the extraction of actionable intelligence from the data more challenging. This methodology-oriented paper focuses on a set of techniques that, when applied to such DNS queries, have proven to help reveal insightful trends and will hopefully lower the barrier of entry to other DNS query investigators.

1. INTRODUCTION

Many commercial and open source tools or applications are widely available to capture, parse and analyze Internet traffic data such as HTTP. Unfortunately, such tools are not as common or prolific for DNS traffic data. Unless one is a subject matter expert in DNS, understanding the nuances of the protocol and techniques to analyze such data can be difficult to achieve and master. This paper focuses on a set of techniques that can be applied to DNS traffic data for the purposes of gaining a deeper understanding of DNS traffic's intent and context. Specifically, we will focus on two main areas of analysis: 1) The query name and 2) The querying source. While there are many other pieces of information stored within the DNS query, the generic techniques described herein can be applied to DNS queries faceted by various data aspects, such as specific DNS header fields.

Thus far, much of the work studying DNS lookup data, particularly as it pertains to the topic of name collisions, has been focused on the relationship between second level domains and top level domains (Interisle Consulting Group, 2013), (JAS/simMachines, 2013). Our goal in this paper is to present techniques and methods, not necessarily findings, that can aide in further dissecting authoritative DNS data for the purpose of better understanding origin and risk associated with it. Furthermore, the techniques described within are not limited to root DNS traffic and can be utilized at the various levels within the DNS hierarchy.

2. TECHNIQUES FOR DOMAIN NAMES

2.1 LABEL SPLITTING

The domain or name being queried within the DNS message is one of the most insightful pieces of data contained within a DNS query. Typically, the name encodes some type of meaningful description of a specific resource or endpoint, e.g. "mail.acme.tld" – which is clearly a mail server within the ACME organization or entity. Parsing and extracting commonalities of various labels or substrings within the queries can help us understand patterns within the requests. The deconstruction of the name is typically done in one of two ways: Label splitting or N-Gram tokenization.

A domain name consists of one or more parts, usually referenced as labels. When represented in text these labels are typically separated by dots to make the domain name easier to read. While it is possible for individual labels to contain dots as a generalization, we will assume throughout this paper that any dot is a label separator. This assumption is made due to some of the data capturing techniques in use on a subset of the name servers that do not preserve or escape the original label separators and instead replace them with dots. Label splitting is a simple technique in which we deconstruct the name into individual terms – one for each label. Simple frequency analysis of labels occurring at the first, second, etc. label is an excellent initial analysis procedure. However, labels at various depths within the name do not always co-align, due to the hierarchical structure in which any entity may choose to encode or represent their endpoints – e.g. "mail.tld" vs. 'mail.sub.tld". Therefore it may also be prudent to conduct label frequency analysis independent of the labels position.

Figure 1 illustrates the distribution of label lengths over a set of DNS queries captured for a single exemplary top-level domain from the A+J root servers in August 2013. Clearly, this abnormal distribution indicates there are areas of interest in which further analysis of the labels at various lengths may prove prudent. In fact, using label splitting frequency analysis over the large corpus of NXDomain names observed at the DNS roots resulted in the discovery of various protocols issuing DNS queries.

Figure 1: Analysis of number of label counts in traffic for one exemplary undelegated TLD

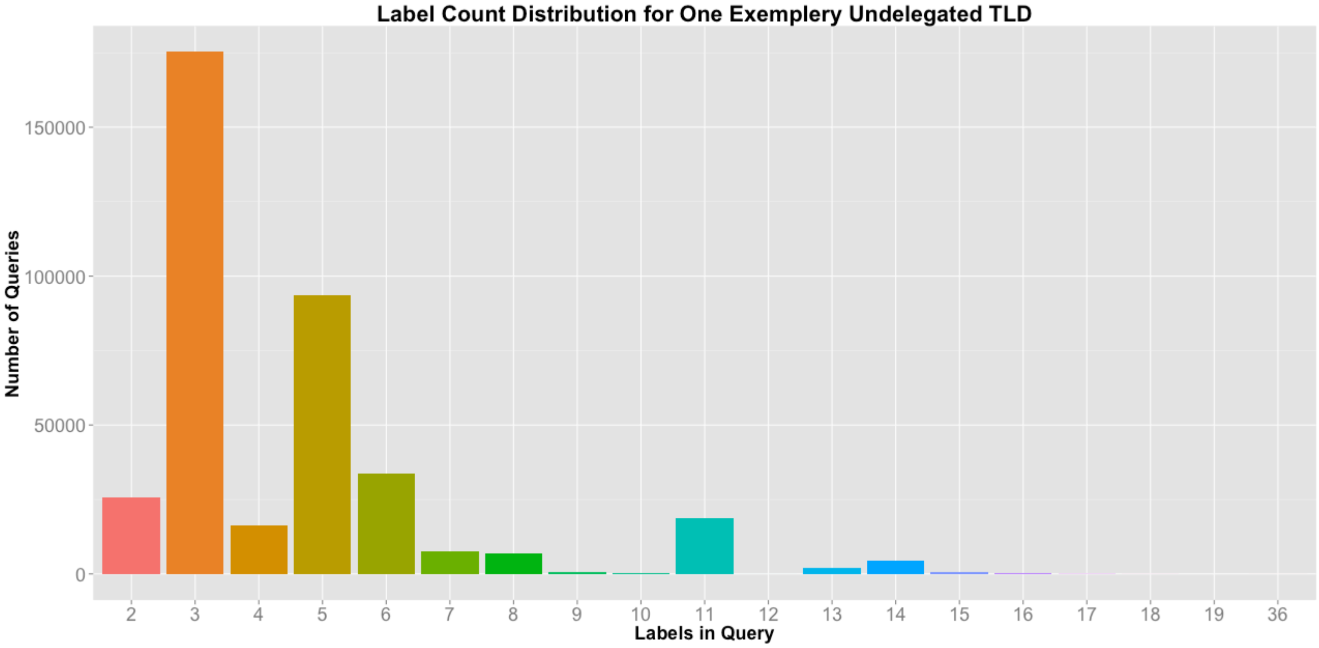


Table 1 outlines the top labels appearing in a sample collection or root data regardless of position in the overall query. A simple yet powerful technique allowed us to discover a multitude of service specific generated names. This data allows us to appropriately tie such services to a name space observed "further to the right" in the domain name. Furthermore, the identified protocol label can be assessed to understand the intent and associated risk to a specific namespace. Many of these service specific protocols were identified in earlier reports (Interisle Consulting Group, 2013) (Verisign Labs, 2013).

Table 1: Most commonly observed labels regardless of position in overall query

Label	
com	home
_tcp	_dns-sd
_msdcs	st
dc	corp
_ldap	wpad
ent	_udp
_sites	us
net	www

2.2 N-GRAM PARSING

An alternative approach to label splitting is N-gram frequency analysis. N-grams are a contiguous sequence of "n" characters from a given sequence of text and are commonly used in probability, communication theory and computational linguistics. Table 2 below shows the decomposition of a domain name into various N-gram lengths. Using N-grams will provide a mechanism to detect commonalities or re-occurring patterns within a set of labels. Table 2 provides an N-gram decomposition of the domain "mail.server.acme.tld".

Table 2: N-Gram decomposition of a domain.

N-Gram Size	N-Grams
1 - Unigrams	mail , server , acme , tld
2 – Bigrams	mail.server , server.acme , acme.tld
3 – Trigrams	mail.server.acme , server.acme.tld

This technique, while more robust than simple label splitting, introduces a computational complexity as N-gram decomposition of a string is $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ permutations, where n is the number of labels within the domain. Furthermore, interpretation of the N-gram frequency list may prove to be challenging, as many of the substrings will not immediately emerge to the top – thus requiring some manual inspection of the results. In addition to N-gram decomposition using labels as a delimiter, the character-by-character decomposition may also be used to search for common substrings within a set of labels.

Such an approach, when applied to the NXDomain names observed at the A and J root servers, revealed the recurring substrings shown in Table 3 below, which isolate the pairings associated with two of the most common labels "_udp" and "_tcp" respectively. These tables help make start to make it clearer that DNS Service Discovery protocols like Bonjour and Microsoft Active Directory lookups could be behind at least a portion of the observed traffic.

Table 2: Common label pairings in root NXD.

Label Combos with TCP	Label Combos with UDP
_ldap._tcp	_dns-sd._udp
_tcp._msdcs	_udp.0
_tcp.dc	_udp.in-addr
_tcp._sites	_udp.arpa
_tcp.cbadomain	lb._udp
_tcp.default-first-site-name	b._udp
_tcp.gc	r._udp
_kerberos._tcp	dr._udp
_tcp.domains	db._udp
_tcp.w-g-c-2	_udp.168

Without the use of N-gram analysis, discovering connected or associated patterns such as the ones in Table 2 would present more of a challenge. This technique provided us a way understanding of the protocols usage patterns. This technique may also be useful for finding organizational structures towards the root of the domain, e.g. "mail.server.department.region.company.tld", in which departments and regions may vary but belong under the same company. One example of this was in the discovery of McAfee GTI protocol activity that was observed during an in-depth analysis of the .CBA top-level domain. GTI clients emit DNS queries whenever files (exe's, pdfs, apks, etc.) are being checked for malware, essentially piggybacking on the DNS. In some circumstances machines were appending an organization specific string at the end of these protocol queries. Two such examples from the earlier presentation were 9.y-0.<label>.<label>.157c.1beb.3ea1.210.0.<label>.avts.mcafee.com.winsinage2.cba and 9.y-0.<label>.<label>.157c.1beb.3ea1.210.0.<label>.avts.mcafee.com.winsinage2.cba (Verisign Labs, 2013). In the labels above the portion of the string that comes after mcafee.com is the portion that identifies the string associated with the organization that most likely has the systems issuing the queries.

2.3 RESPONSIBLE LABEL ANALYSIS

During the analysis of domain names it is possible that personal identifiable information and/or corporate information may be revealed, and appropriate handling and reporting of that data takes precedent. Techniques such as label splitting or N-gram analysis has proven to unveil data such as internal hostnames of corporate networks, IP addresses of specific entities and hashes of data or resources within the root NXDomain name traffic data.

3. TECHNIQUES FOR QUERYING SOURCE ADDRESS

3.1 IP ADDRESS ABSTRACTION AND MAPPING

The querying source information of DNS request can be utilized to give more color to a specific name space as it describes the resource from which it is being requested. The combination of the querying IP address and time a query is made can be used to gauge traffic diversity measurements, geographical affinities and longitudinal patterns. In order to measure these trends, an IP address will be required to map into various forms of network and geographic representations, such as /24 netblocks, autonomous systems and geographical locations such as a country or state.

DNS queries for a namespace may span a multitude of unique querying source IP addresses, which makes analyzing query traffic patterns more difficult. In order to overcome this challenge, mapping a specific IP address for a given query to a more generic and broad range or identifier is typically suitable. IP addresses can be designated in many various network allocation ranges; however, it is common for an entity such as company or ISP to own or utilize a /24 network range. Mapping a given querying IP source to a /24 netblock, as well as the containing autonomous system, is usually sufficient for grouping and frequency analysis of namespaces.

Table 3: IP Address Abstraction of Specific TLDs Measured at A & J Roots.

TLD	Total Requests	Unique IPs	Unique /24	Unique ASNs
HOME.	**2727531510**	**481568**	**302307**	**23305**
CORP.	**404853888**	**261393**	**171728**	**19672**
BOX.	**33585163**	**258354**	**128588**	**9876**
MAIL.	**18391999**	**425019**	**279863**	**19838**
LIVE.	**2311797**	**103354**	**78480**	**8645**

This decomposition of network traffic for a given domain or subdomain provides a convenient way to gauge the diversity of traffic. One can imagine large corporations that operate over a vast IP range could produce many common DNS queries from a variety of unique IP addresses; however, when abstracted to a higher entity structure such as ASNs, the traffic could more easily be attributed to that company empirically. Table 3 depicts IP address decomposition for various TLDs measured from the A and J root servers, in which the ratio of unique querying IPs to /24 netblocks and ASNs clearly varies between different TLDs.

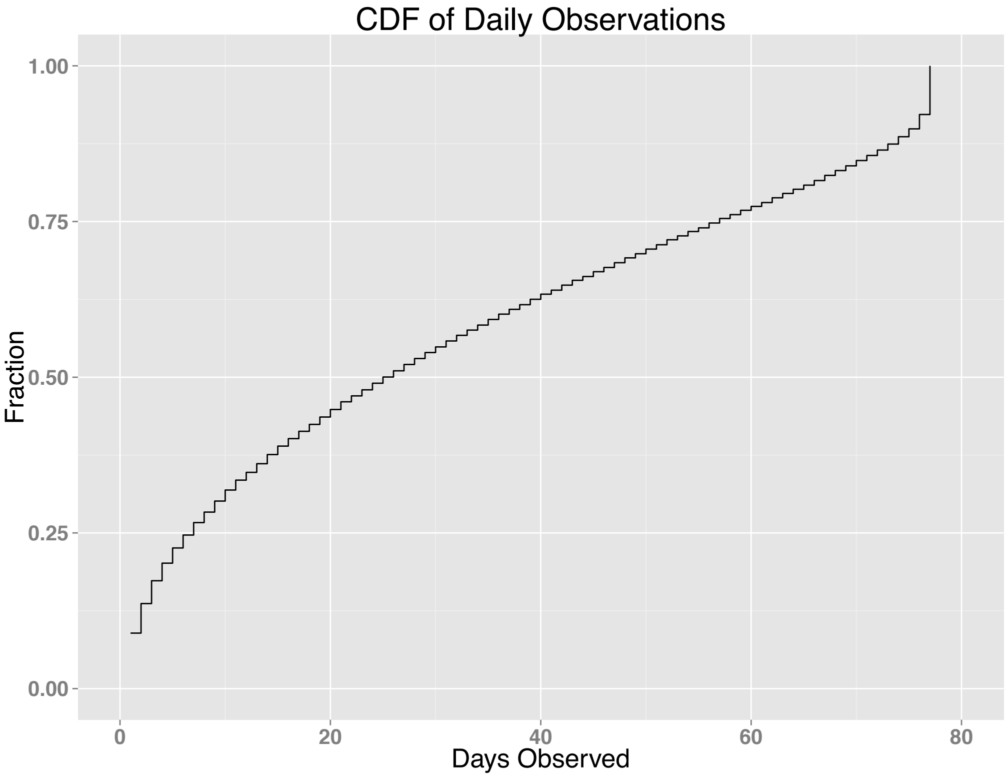
These various traffic diversity measurements have been used to gauge levels of associated risk – e.g. ".home" and ".corp" (Verisign Labs, 2013). Understanding the distribution of traffic diversity for a given namespace may allow the investigator to prioritize or limit their next analysis measurements. However, studying traffic diversity measurements over a longitudinal period of time may reveal diverging trends.

3.2 LONGITUDINAL AND PERIODICITY MEASUREMENTS

Measuring traffic volume and diversity measurements at various network allocations over time for a specific name or substring can also provide insightful information. Many of the requests under a specific name space may in fact be a singleton event - as the string in question is a highly entropic entity. A common example of such a string is the ten-character name generated by the Chrome browser at startup and is used to discover prosperities of network behavior (Google Product Forums, 2010). Measuring time distributions and request frequencies from both a network and name point-of-view can be helpful in discerning machine generated from that of systematic and or human generated traffic.

Such a simple technique would be to count the number of time periods a specific namespace appears in the given collection window. One might, for example, count the number of days a namespace appeared within a week or month. Figure 3 is an exemplary Cumulative Distribution Function (CDF) of such a set of measurements. The figure reflects number of days a specific domain within a TLD was observed over a 85 day collection period from the A and J root servers. The upper and lower bounds of the CDF distribution have proven useful for identifying singleton contributors and those with a more systematic or high frequency pattern. Alternatively, more narrow distributions can be achieved by looking at a specific CDF distribution of a domain or subdomain associated with a particular network entity such as a netblock or ASN. Overall, this technique is useful for identifying the entities with extreme periodicity measurements; however, the majority of namespaces not at the extremities of the distribution will require an additional analytical approach to understand their request patterns.

Figure 2: An exemplary CDF of daily observations of domains within a TLD.



For non-singleton entities the distribution of inter-query times can be used to understand the periodicity of the entity. Equation 1 provides a measurement of the average inter-query rates for a particular namespace. For a given entity, the sequential time deltas between queries forms a distribution of inter-query time intervals in which useful statistical measurements, such as the mean, as depicted in Equation 3, can be calculated. Other statistical measurements over the distribution such as the maximum inter-query time, as shown in Figure 3, may also prove useful for understanding the domains requests pattern.

The measurements of inter-query time may unveil the "burstiness" or "rhythmic" request pattern for a given namespace from a specific network. Regular rhythmic traffic patterns for specific namespaces may be an early indicator of well established internal usage of a non-delegated namespace while sporadic requests patterns may require additional analysis techniques.

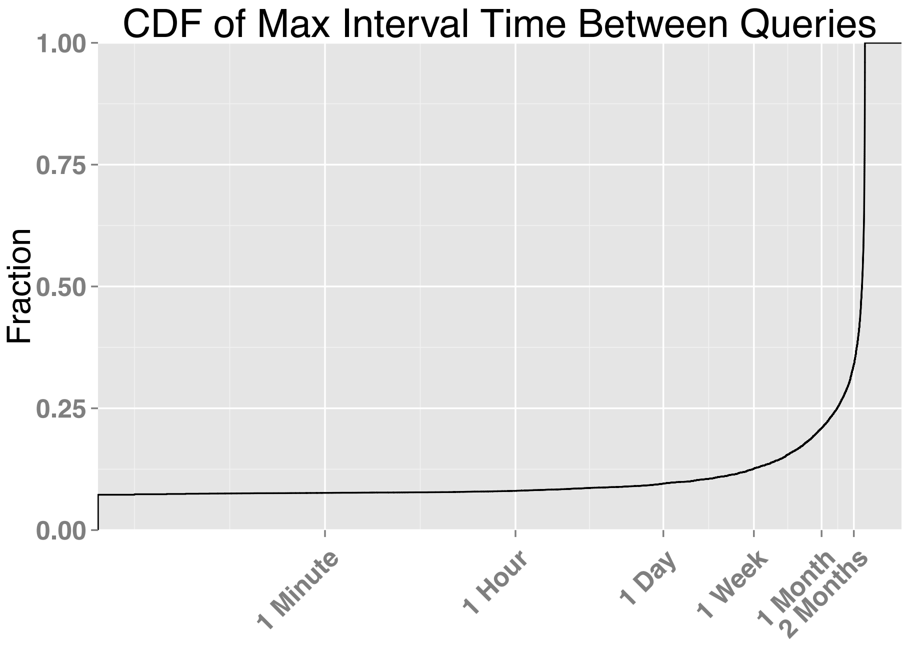
Equation 1: Measuring inter-query time intervals for a namespace.

$$\Delta_{ki} = \tau_i(\varepsilon_k) - \tau_{i-1}(\varepsilon_k)$$

$$\mu_k = \frac{\sum_{i=1}^{n} \Delta_{ki}}{n}$$

ε_k : measured domain
τ_i : time of measured request
τ_{i-1} : time of last measured request

Figure 3: An exemplary CDF of max inter-query time intervals for domains.



3.3 REGIONAL AFFINITY

Geographical mappings of IPs are only reliable as the underlying geo-to-IP data set. Most of the commercially available geo-to-IP datasets are focused on end client IP address mappings and not Internet infrastructure IP mappings, which most of the DNS queries presumably originate from. Therefore, with limited accuracy and known false-positives rates of geo-to-IP data, mapping an IP address to a more generic and broad geographical region such as a country is most likely more reliable than a specific street or city mapping. (Wikipedia, 2014)

For doing broad scale analysis of where certain strings are preferred we have developed a technique to analyze the source IP addresses for a given set of strings to find out if any regions have a particular affinity for one of the strings that is disproportionate to its affinity for the other strings in the set. By identifying the specific countries that have an affinity for an applied for string, it is easier to further investigate what is generating these queries for the purpose of risk analysis.

The equation that we have derived normalizes the amount of traffic any string gets from each country by the total amount of traffic from that country. Now, with the normalized traffic value for a string from a given country it is possible to determine the average percentage any country is expected to have for a given string, as is shown in Equation 2 below. For any country, the standard deviation across all of the strings it requested can be calculated and the number of standard deviations above average that a given country has for a specific string can be computed. The results of this process are illustrated in Equation 4 and the raw data is available at:

http://www.verisignlabs.com/documents/Verisign%20Applied%20for%20String%20Regional%20Affinity.xlsx.

Equation 2: Determining regional affinity first requires normalizing traffic for a string across all of the countries requesting the string

$$i_c^{AFS} = \frac{q_c^{AFS}}{Q_c}$$

$c = country$

$AFS = Applied\ for\ String$

$i_c^{AFS} = Proportion\ of\ queries\ for\ AFS\ from\ c$

$q_c^{AFS} = Number\ of\ queries\ for\ AFS\ from\ c$

$Q_c \ = Total\ queries\ from\ a\ c$

Equation 3: Using the normalized scores from Equation 1 we can establish a baseline for the normal level of interest any string should receive from a particular country

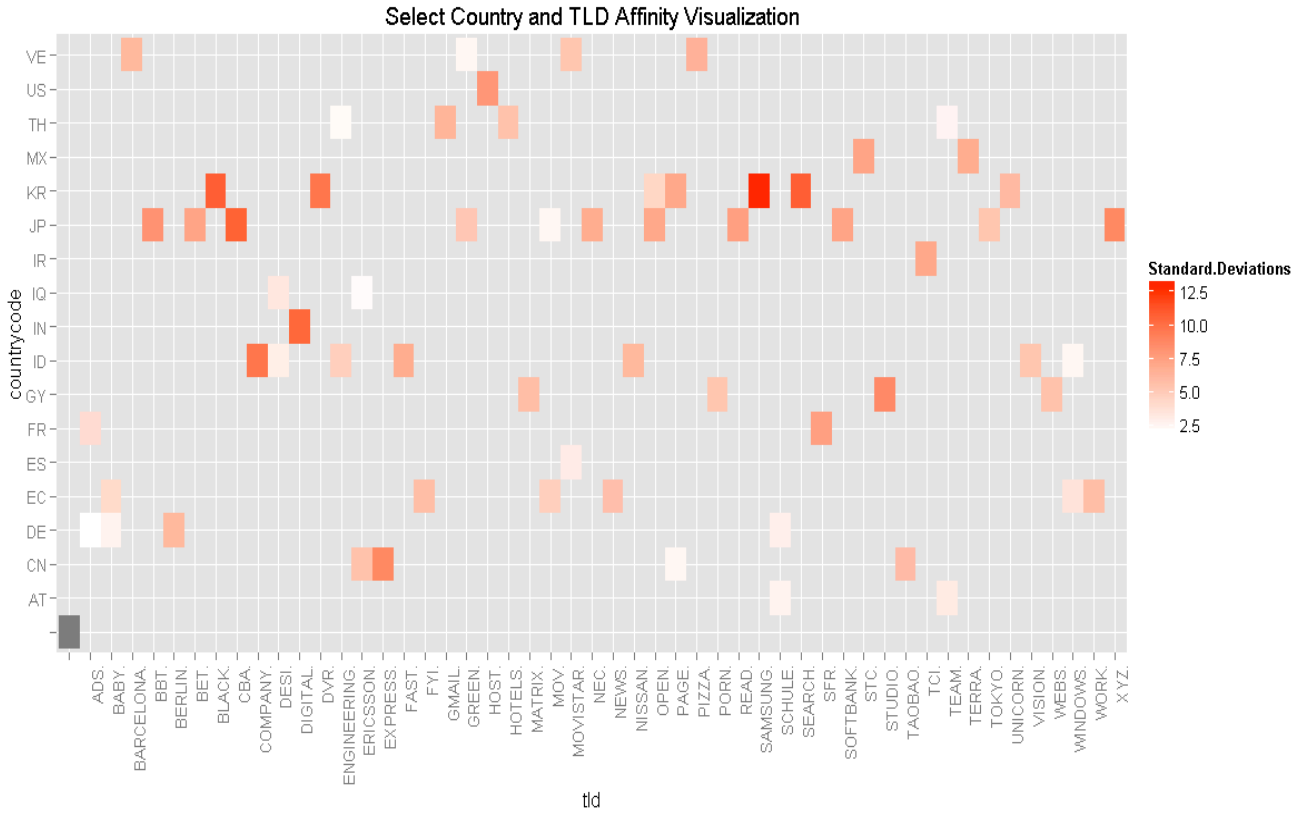
$$I^{AFS} = \frac{\sum_{c=1}^{N} i_c^{AFS}}{N}$$

$I^{AFS} = \ Average\ of\ Country\ Percentages\ for\ an\ AFS$

$N = \ Number\ of\ Countries\ that\ meet\ minimum\ traffic\ threshold$

$i_c^{AFS} = Proportion\ of\ queries\ for\ AFS\ from\ a\ country$

Figure 4: Selected results from the affinity calculations show how certain strings are preferred by some regions



Conclusion

This paper illustrated numerous ways in which DNS data can be analyzed to understand the contextual intent of queries. We first illustrated how domain names can be deconstructed to discover commonalities and patterns within various labels. Next we extended those techniques to identify distinct namespaces that are likely to correlate with impacted user bases and attributed to specific protocols. Finally, we illustrated techniques for leveraging ASN and querying source IP information to better understand the observed DNS traffic data and monitoring the longitudinal query rates of various labels. While the techniques and methods described in this paper are not exhaustive, we believe this general framework can be used as a guideline for analyzing and understanding DNS traffic patterns in general, and will be helpful in diagnosing potential name collision possibilities. While outside of the scope of this paper, we believe the techniques described within could be used as a foundation of a DNS data framework for analyzing packet captures of authoritative traffic and generate actionable reports. Ideally, as more individuals become interested in understanding the traffic that has been observed regarding strings they have applied for, or begin collecting their own traffic, they can apply these techniques on their own to better understand the traffic that they will be responsible for.

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.

4. Works Cited

Google Product Forums. (2010, 06 12). *Chrome causes six different DNS requests for random 10 character host names on startup*. Retrieved 02 26, 2014, from Google Product Forms: https://productforums.google.com/forum/#!topic/chrome/dQ92XhrDjfk

Interisle Consulting Group. (2013). *Name Collision in the DNS.*

JAS/simMachines. (2013, 09 17). *Namespace Expansion.* Retrieved 02 07, 2014, from JAS Advisors: https://www.jasadvisors.com/namespace-expansion-i.pdf

Verisign Labs. (2013, 09 15). *ICANN's Proposal to Mitigate Name Collision Risks – .CBA Case Study.* Retrieved 02 26, 2014, from Verisign Inc.: https://www.verisigninc.com/assets/report-cba-analysis.pdf

Verisign Labs. (2013, 08 22). *New gTLD Security, Stability, Resiliency Update: Exploratory.* Retrieved 02 25, 2014, from Verisign Labs Technical Reports: http://techreports.verisignlabs.com/docs/tr-1130008-1.pdf

Wikipedia. (2014, 02 21). *Geolocation Software*. Retrieved 02 26, 2014, from Wikipedia: http://en.wikipedia.org/wiki/Geolocation_software