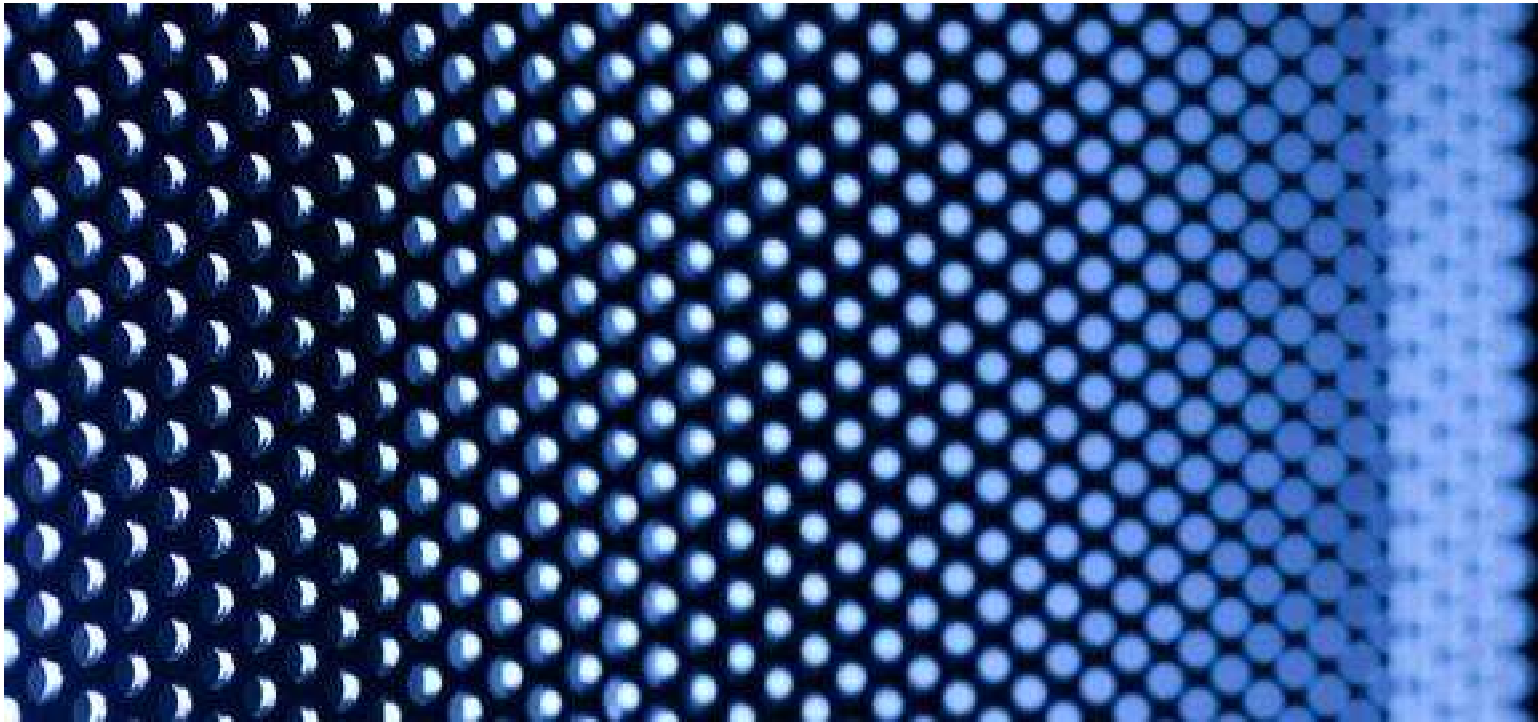


Analysis Techniques for Determining Cause and Ownership of DNS Queries

Andrew Simpson

Matthew Thomas

March 2014

About this talk

- A methodology-oriented presentation focused on DNS analysis techniques for measuring and understanding cause and ownership of queries.

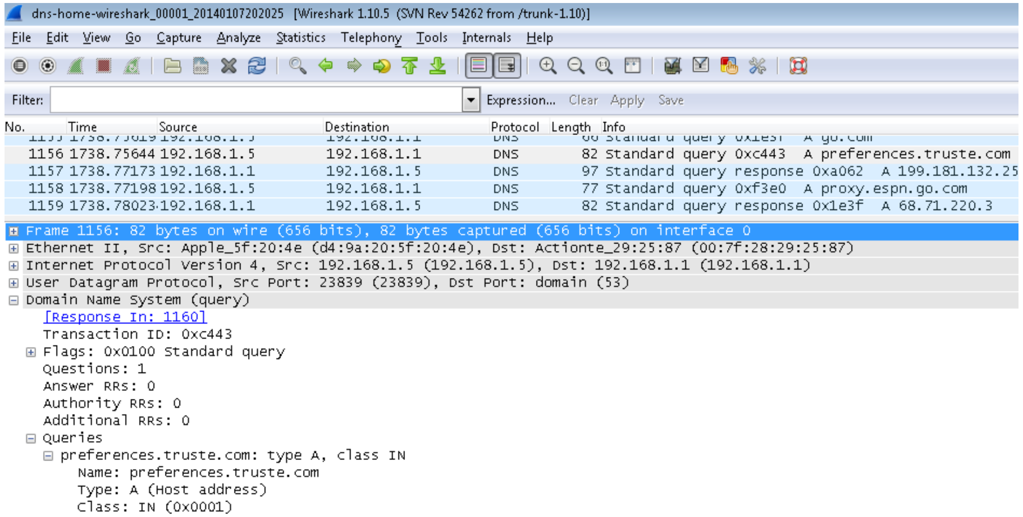
No Turnkey Solution for DNS Analysis

Getting Ahold of DNS Data

- Existing Repositories
  - “Day-in-the-Life” or DITL
    - Originally conceived by CAIDA to help DNS root operators study and improve the integrity of the root server system
    - Maintained by DNS Operation and Research Center (DNS-OARC)
    - Consists of DNS query and responses over a continuous 48-hour sample from various Root DNS operators for the past several years
- Capture Your Own
  - DNS-OARC tools to instrument your own network
    - https://www.dns-oarc.net/tools/dnscap
    - Network capture utility designed specifically for DNS traffic. It produces binary data in pcap(3) format.

Analyzing PCAP

- Most DNS collection results in packet capture (PCAP) files
  - Browsing PCAP with graphical utility like Wireshark



  - Scaling to TB size datasets
    - https://github.com/packetloop/packetpig
    - An Open Source Big Data Security Analytics tool that analyses pcap files using Apache Pig.

Foundational Elements of a DNS Analysis Framework

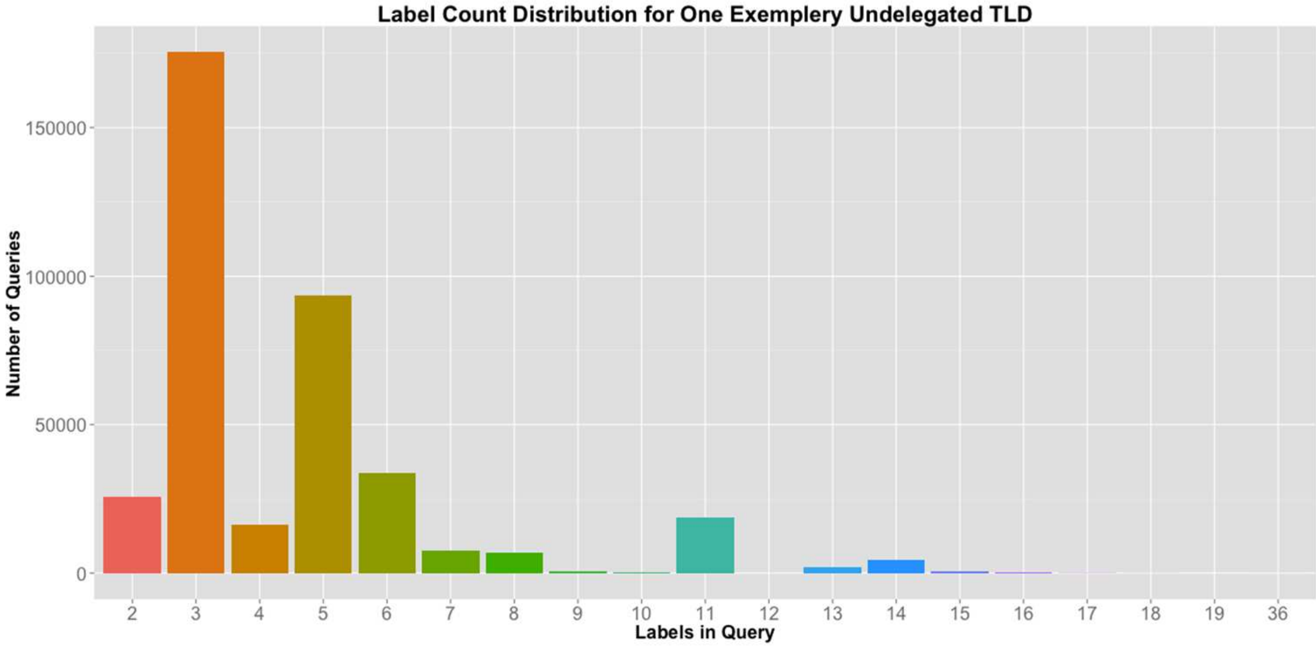
- Domain Name Decomposition
  - Protocols
  - Organization/Entity Structure
- Time of Request
  - Periodicity
  - Traffic Patterns
- Requesting Source IP
  - Diversity Measurements
  - Geographical Affinities

Faceted Analysis

- Techniques are generic and can be applied at various levels within the DNS hierarchy
- Applied to specific facets of the DNS query or aspects of a particular data range.

Domain Name Analysis

- Typically the domain name encodes some type of meaningful description of a specific resource
  - mail.acme.tld
- Deconstructing the name into individual labels can be achieved by splitting the domain by the dot delimiter



Label Decomposition

- Specific label "depth" analysis (e.g. first, second, third)
  - Entity / Organization predominance

- "Depth" Agnostic analysis
  - Protocol identification

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

N-Gram Decomposition

- More robust alternative approach to individual label splitting.
- N-Grams: contiguous sequence of "n" characters from a given sequence of text.

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

- Introduces a computational complexity $$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$
- N-Gram at character level instead of labels
  - Common sub-strings within labels

N-Gram Decomposition

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

Information Exchange and Reporting

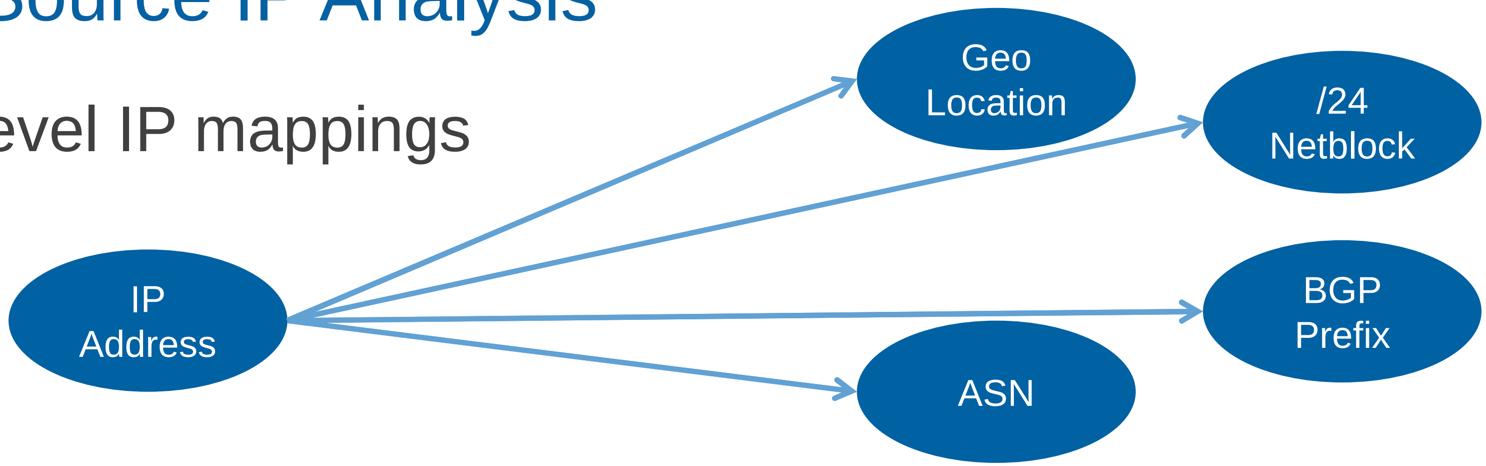
…

9.y-0.**<label>.<label>**.157c.1beb.3ea1.210.0.**<label>**.avts.mcafee.com.winsinage2.cba 9.y-0.**<label>.<label>**.157c.1beb.3ea1.210.0.**<label>**.avts.mcafee.com.winsinage2.cba

…

Query Source IP Analysis

- Higher level IP mappings



- Useful for determining reach / diversity of traffic.

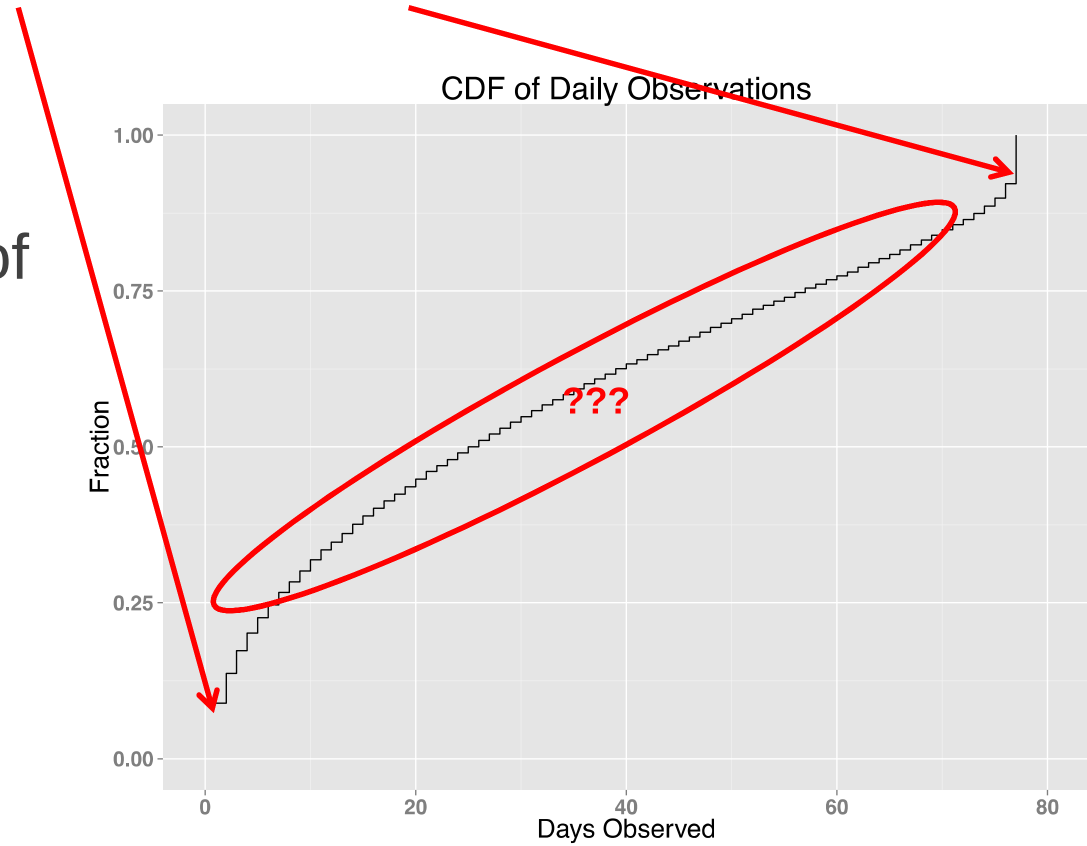
TLD	Total Requests	Unique IPs	Unique /24	Unique ASNs
HOME.	2727531510	481568	302307	23305
CORP.	404853888	261393	171728	19672
BOX.	33585163	258354	128588	9876
MAIL.	18391999	425019	279863	19838
LIVE.	2311797	103354	78480	8645

- Special care should be given to geo-IP
  - Infrastructure IP space is typically not as accurate as end-user.

Frequency Analysis

- Measuring occurrence rates can identify trends or general areas of interest
  - Singleton events (Chrome NXDs) vs. Persistent

- Exemplary CDF plot measuring the number of days a domain within a TLD was observed over an 85 day collection period from A & J.



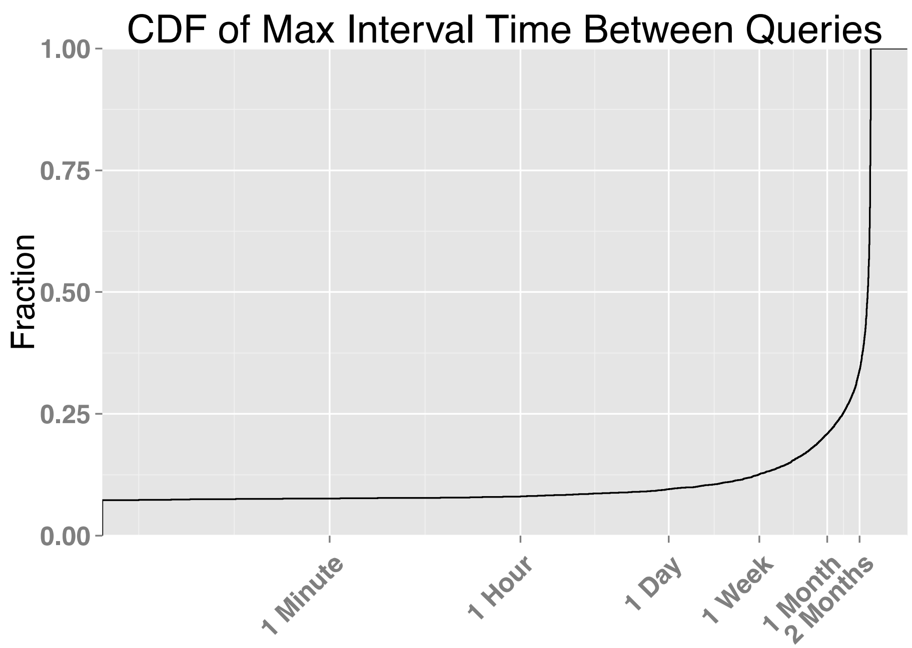
Periodicity

- What is the average periodicity of a domain's requests?
- Measure time between sequential requests.
- Calculate other statistical measurements on distribution.

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

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

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



Geographical Affinities

- DNS Queries for some applied for strings originate disproportionately from certain countries
  - Root server data currently allows us to study queries for the more than 1,400 applied for strings with an NXDomain response

- The outlined method can be applied to captured data for any set of strings a server is authoritative for
  - By identifying the specific countries that have affinity for an applied for string, it is easier to further investigate what is generating these queries for the purpose of risk analysis
  - If performing this analysis at an authoritative level below root, it is possible to further segment affinity by second level domain or lower

Regional Data Assignment

- Destination IP Augmented with 2-letter country code using Maxmind GeoIP data
  - Aggregates are generated with raw query count by TLD by country

Applied for String	Country Code	Query Count
newtld1	AE	40
newtld1	AL	16
newtld1	AO	11
newtld1	AR	10
newtld1	AS	1
newtld2	AE	36
newtld2	AL	22
newtld2	AO	13
newtld2	AR	96
newtld2	AS	2

	Origin of Query				
Applied for String	AE	AL	AO	AR	AS
newtld1	40	16	11	10	1
newtld2	36	22	13	96	2
Region Totals	76	38	24	106	3

Normalizing for Regional Preferences

- On average, what proportion of the queries originating from a specific country are resolving a particular applied for string?

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

$c = country$

$AFS = Applied\ for\ String$

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

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

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

- When Q_c is less than .01% of Q (the total observed query count) the queries from that country are not considered to avoid introducing volatility from countries where queries may no

	Origin of Query (c)				
Applied for String	AE	AL	AO	AR	AS
newtld1	40	16	11	10	1
newtld2	36	22	13	96	2
Country Totals(Qc)	76	38	24	106	3

Applied for String	AE	AL	AO	AR	AS
newtld1	52.6%	42.1%	45.8%	9.4%	33.3%
newtld2	47.4%	57.9%	54.2%	90.6%	66.7%

Establishing Baselines for Regional Preference

- The percentages serve as normalized values to compare countries for a given applied for string
  - The baseline for what is expected from a country is the average of all country proportions for an applied for string

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

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

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

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

  - The standard deviation of the proportions for an applied for string are then used to determine how far off the baseline any individual country is

% Distribution by TLD	Origin of Query (c)						
Applied for String	AE	AL	AO	AR	AS	Average	Standard Deviation
Newtld1	52.6%	42.1%	45.8%	9.4%	33.3%	36.7%	15.0%
newtld2	47.4%	57.9%	54.2%	90.6%	66.7%	63.3%	15.0%

Standard Deviations	Origin of Query (c)				
Applied for String	AE	AL	AO	AR	AS
newtld1	1.07	0.36	0.61	-1.82	-0.22
newtld2	-1.07	-0.36	-0.61	**1.82**	0.22

AR has an affinity for newtld2

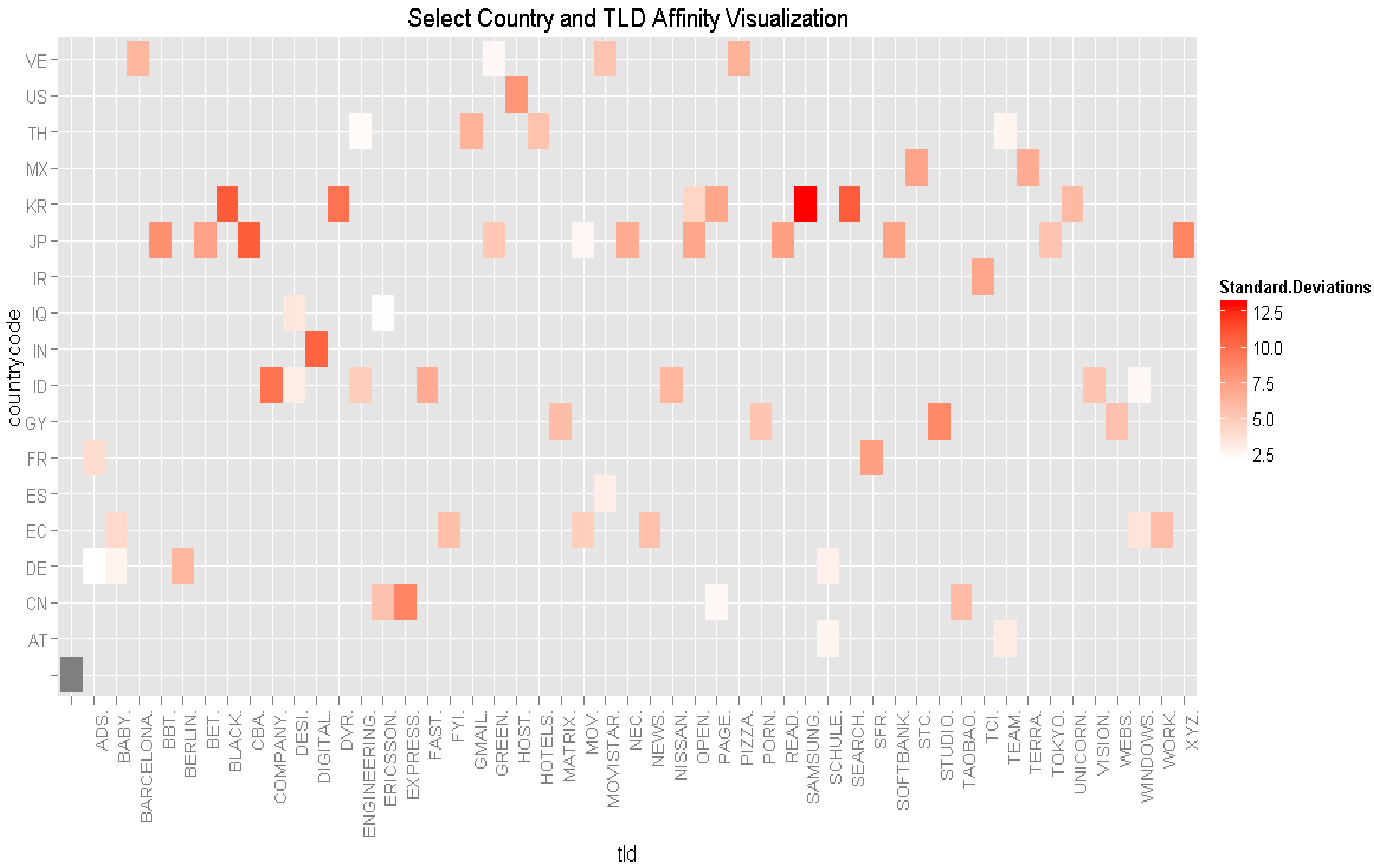
Raw Results

- Subset of full results

Originating Country/ Applied for String	Standard Deviations
DE	
.BERLIN	6.12
.SCHULE	2.86
.BABY	2.63
.COLOGNE	2.17
.HAUS	2.13
JP	
.CBA	10.69
.XYZ	8.85
.BBT	8.20
.READ	7.42
.BET	7.28
US	
.HOST	7.94
.WOW	5.17
.DENTAL	3.29
.COMCAST	2.75
.ANTHEM	2.37

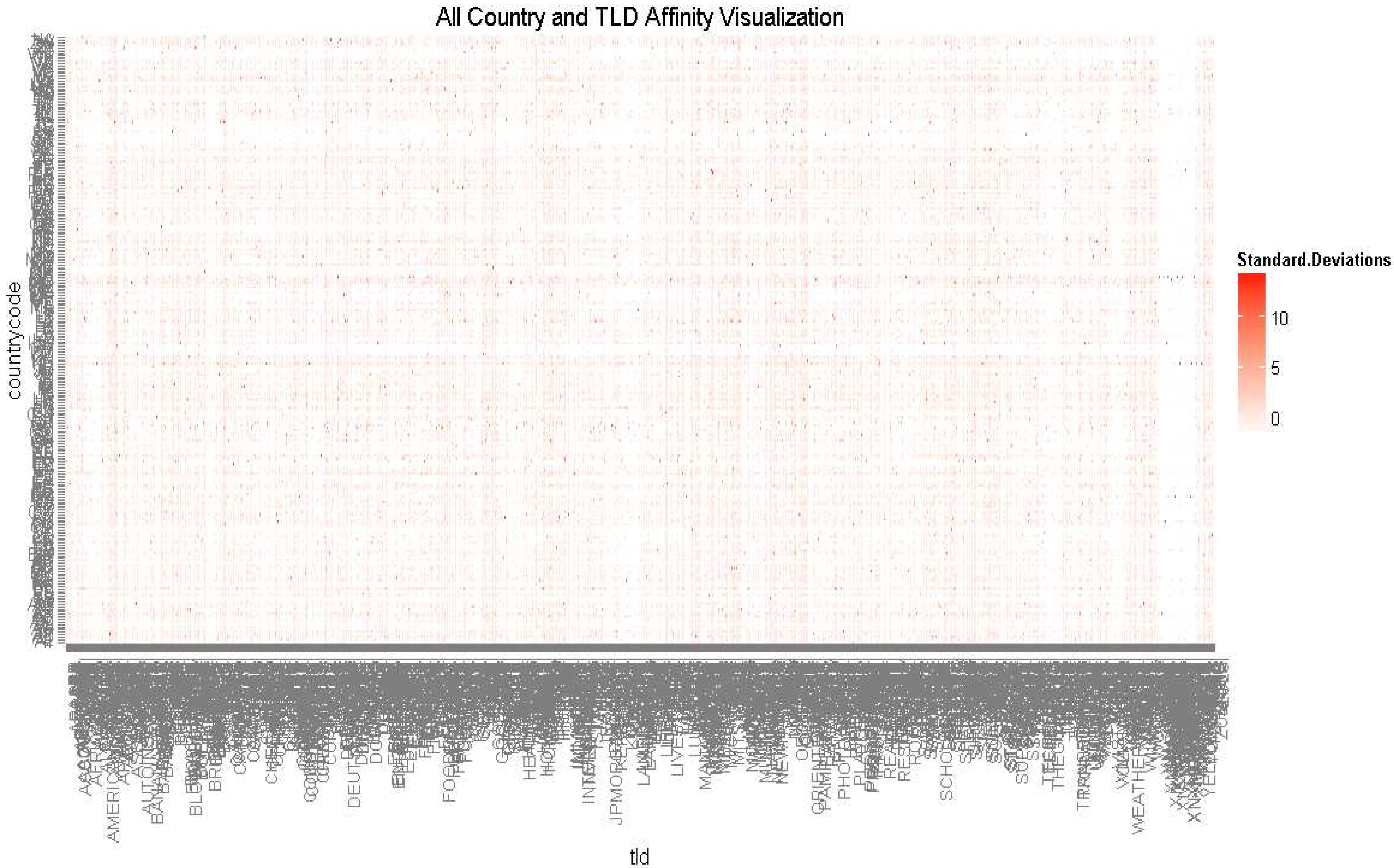
Originating Country/ Applied for String	Standard Deviations
FR	
.SFR	7.44
.BZH	5.05
.LOREAL	4.67
.ADS	3.98
.PROD	3.75
KR	
.SAMSUNG	13.04
.BLACK	10.81
.SEARCH	10.78
.DVR	9.77
.PAGE	7.10
ZA	
.MARRIOTT	4.35
.DURBAN	3.20
.EVENTS	3.19
.SKY	2.98
.CLOUD	2.36

Selected Results



Complete Results Visualization

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



In Summary

- Studying DNS queries can provide insights into their root origin
  - The hostnames encode semantically meaningful details about who is asking for what
  - The source of the queries provides details about the who that can further break down the problem and help scope risk

- Operators are the best equipped to understand their query patterns
  - Those closer to the source of the traffic can ultimately get more data that can better explain the root causes
  - Operators are the best equipped to impact a change in their network configurations

powered by

VERISIGN™