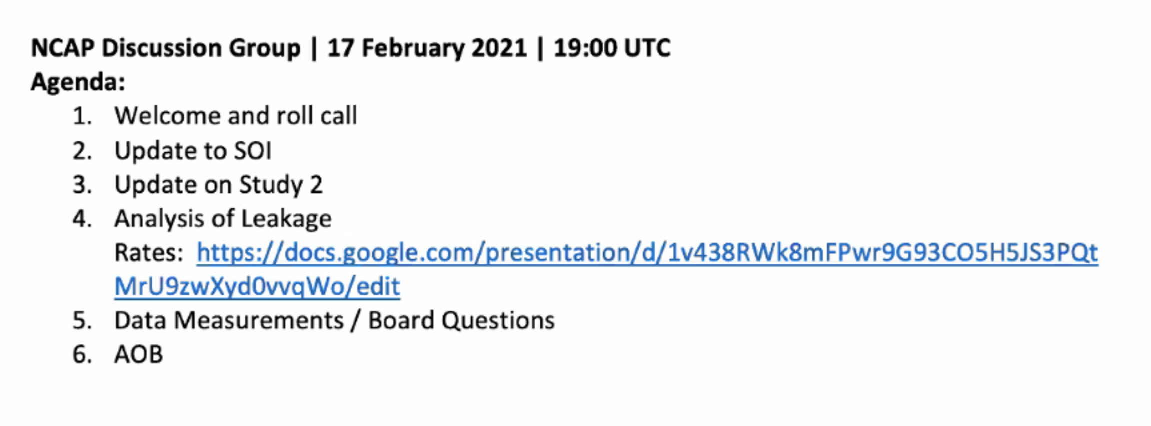


NCAP Discussion Group | 17 February 2021 | 19:00 UTC
Agenda:

1. Welcome and roll call
2. Update to SOI
3. Update on Study 2
4. Analysis of Leakage Rates: https://docs.google.com/presentation/d/1v438RWk8mFPwr9G93CO5H5JS3PQtMrU9zwXyd0vvqWo/edit
5. Data Measurements / Board Questions
6. AOB

Table of Contents

ACTION ITEMS

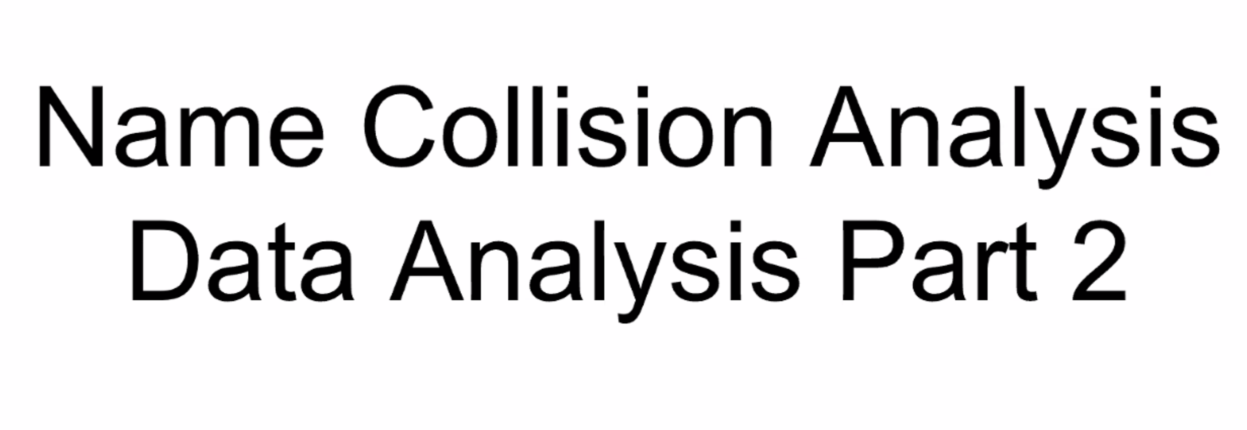
MEETING DATE	DESCRIPTION	OWNERSHIP
17 Feb 2021	Suggestion that a group workspace be utilized to start noting team's thoughts on what measurements to pursue, etc.	Team

	No decision made on 17 Feb; suggest this be re-discussed at next meeting	

SOI: None

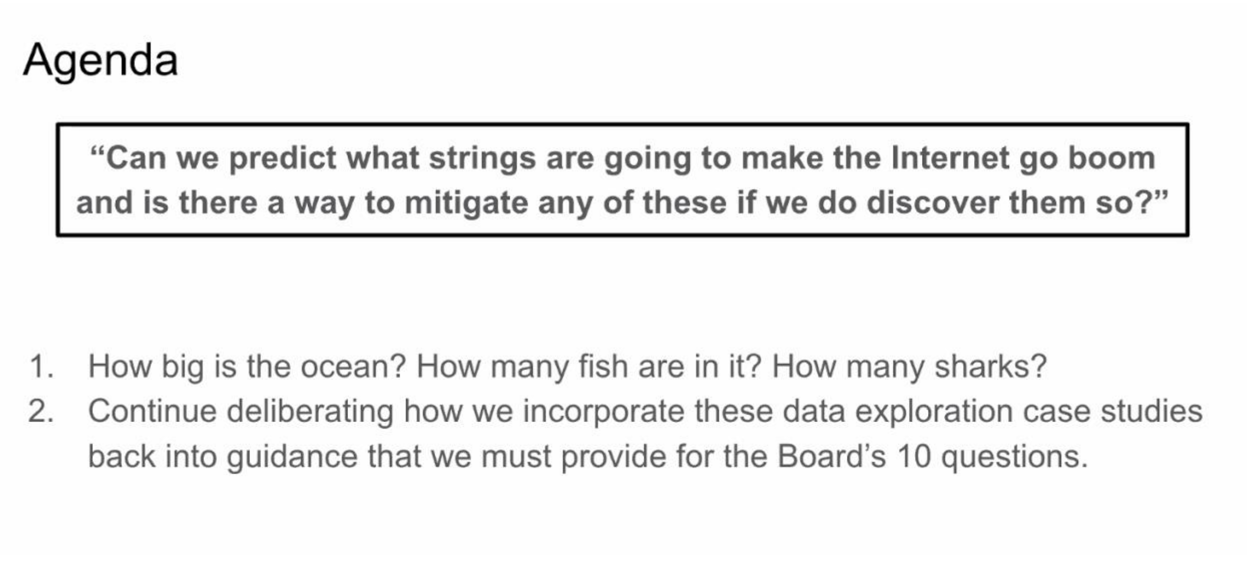
STUDY 2: SSAC sometimes sends docs to ICANN Legal to check it – very appropriate to do so and not giving ICANN unnecessary influence over documents. Study 2: conflict of interest text has been modified so we will send to LEGAL to review again. Next goes to BTC.

PRESENTATION: https://docs.google.com/presentation/d/1v438RWk8mFPwr9G93CO5H5JS3PQtMrU9zwXyd0vvqWo/edit#slide=id.p



Name Collision Analysis
Data Analysis Part 2

SLIDE 1: Agenda



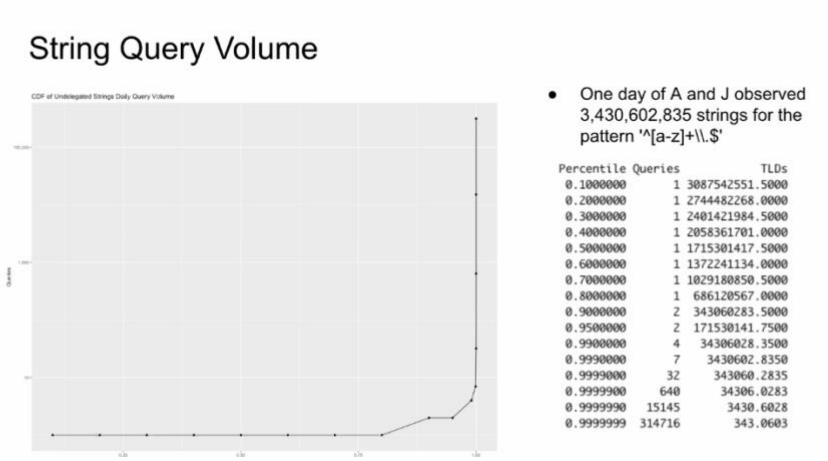
Agenda

"Can we predict what strings are going to make the Internet go boom and is there a way to mitigate any of these if we do discover them so?"

1. How big is the ocean? How many fish are in it? How many sharks?
2. Continue deliberating how we incorporate these data exploration case studies back into guidance that we must provide for the Board's 10 questions.

Want to determine distribution of query strngs

SLIDE 2: String Query Volume



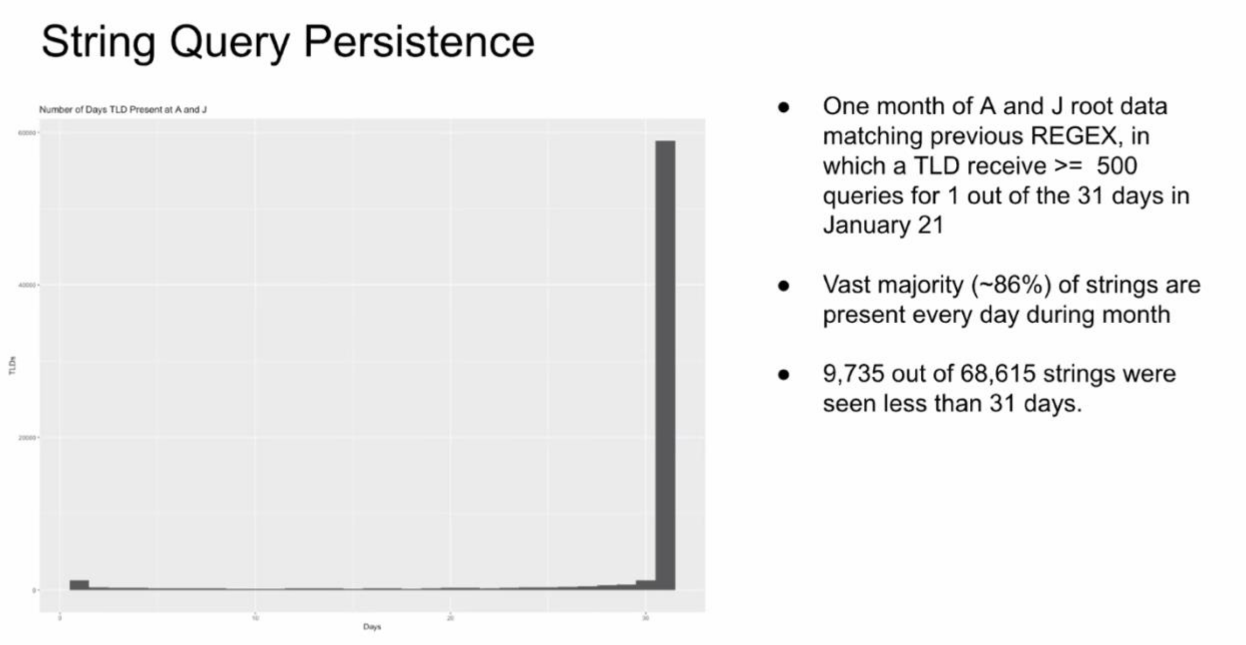
String Query Volume

- One day of A and J observed 3,430,602,835 strings for the pattern '^[a-z]+\\.$'

```
Percentile Queries             TLDs
 0.1000000       1 3087542551.5000
 0.2000000       1 2744482268.0000
 0.3000000       1 2401421984.5000
 0.4000000       1 2058361701.0000
 0.5000000       1 1715301417.5000
 0.6000000       1 1372241134.0000
 0.7000000       1 1029180850.5000
 0.8000000       1  686120567.0000
 0.9000000       2  343060283.5000
 0.9500000       2  171530141.7500
 0.9900000       4   34306028.3500
 0.9990000       7    3430602.8350
 0.9999000      32     343060.2835
 0.9999900     640      34306.0283
 0.9999990   15145       3430.6028
 0.9999999  314716        343.0603
```

1 days worth of queries. String that contains only characters that are a thru z. 3 billion strings that matched that pattern, distinct strings ones. What is distribution of distinct strings look like in terms of query volume. 99% percentile are receiving less than 4 queries. Very limited pool. There is a finite # of strings we are looking at.   Volumetric based strings is important context.

SLIDE 3: String Query Persistence

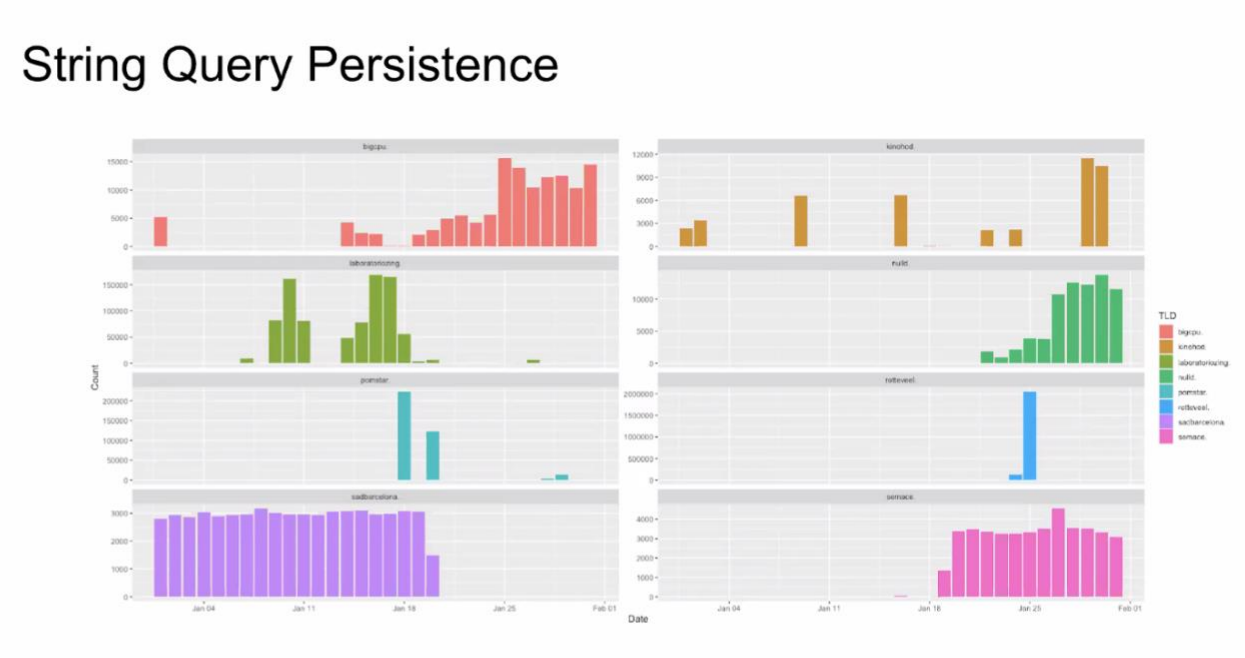


String Query Persistence

- One month of A and J root data matching previous REGEX, in which a TLD receive >= 500 queries for 1 out of the 31 days in January 21
- Vast majority (~86%) of strings are present every day during month
- 9,735 out of 68,615 strings were seen less than 31 days.

Frequency?   Measured # of days we saw. 86% were seen daily.
How long do you need to do a data capture then?

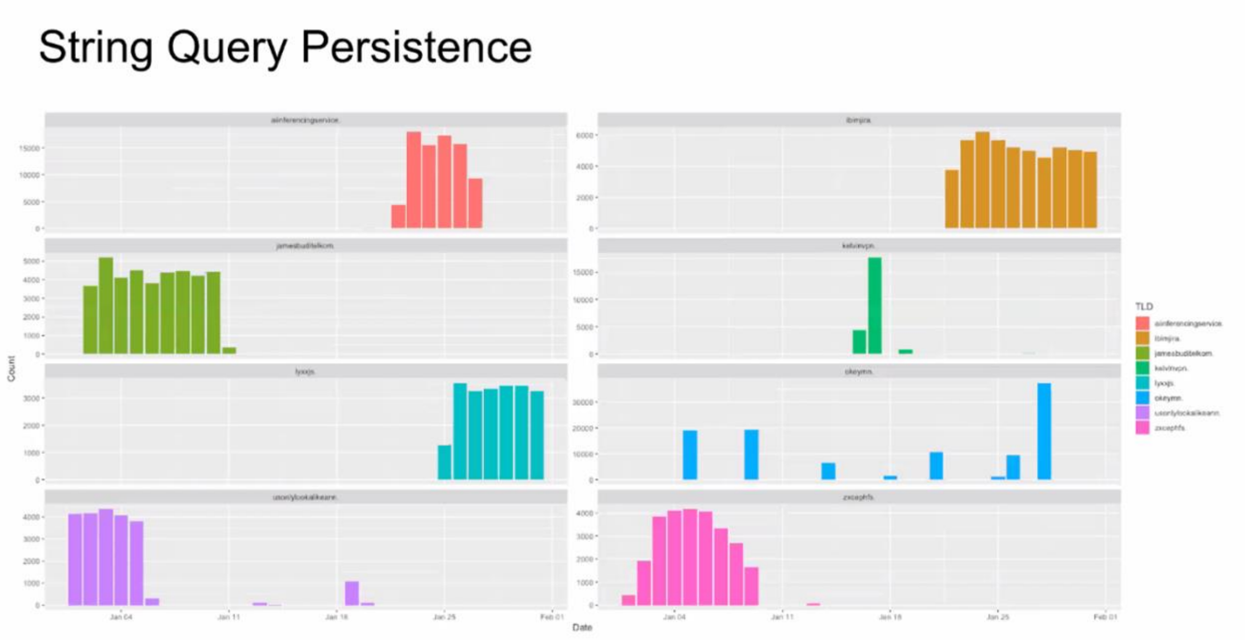
SLIDE 4: String Query Persistence 2



String Query Persistence

strings only present for 13-14 days out of 31 days. Why do some disappear? Do they fix their strings?

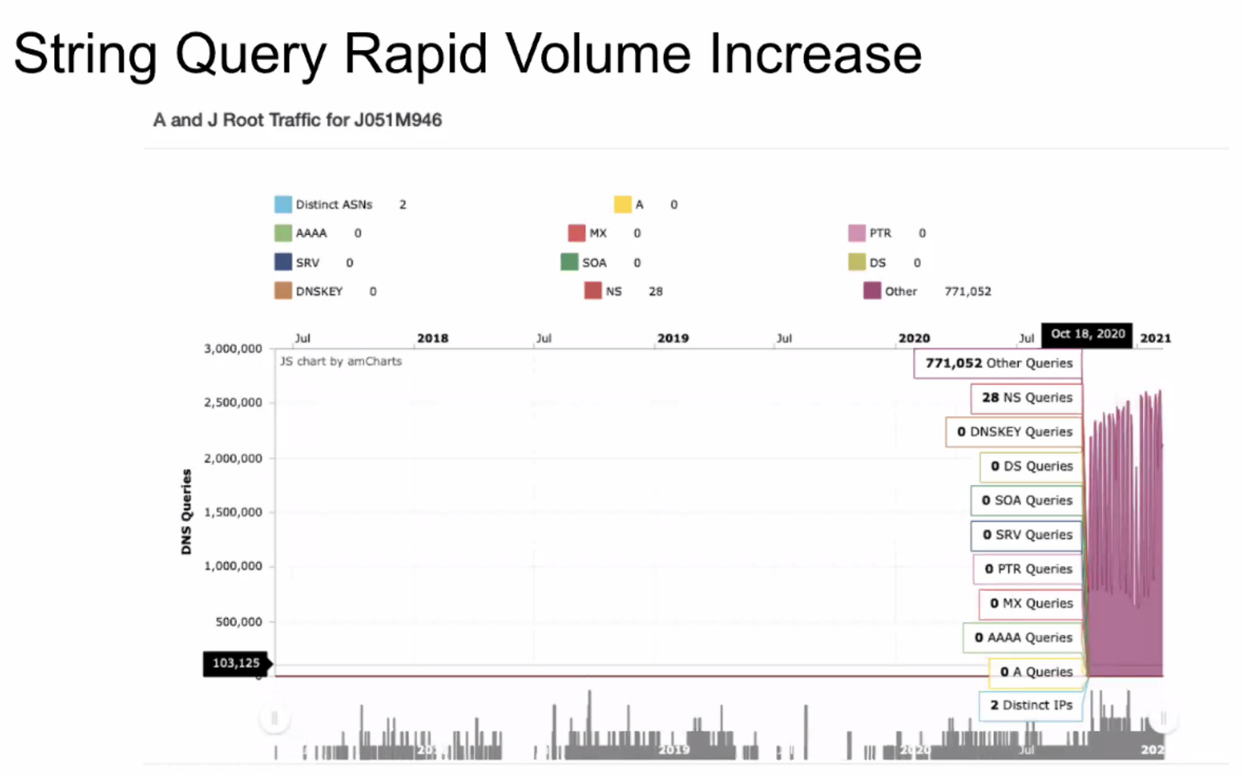
SLIDE 5: String Query Persistence 3



String Query Persistence

THESE were present for 6 or 7 days

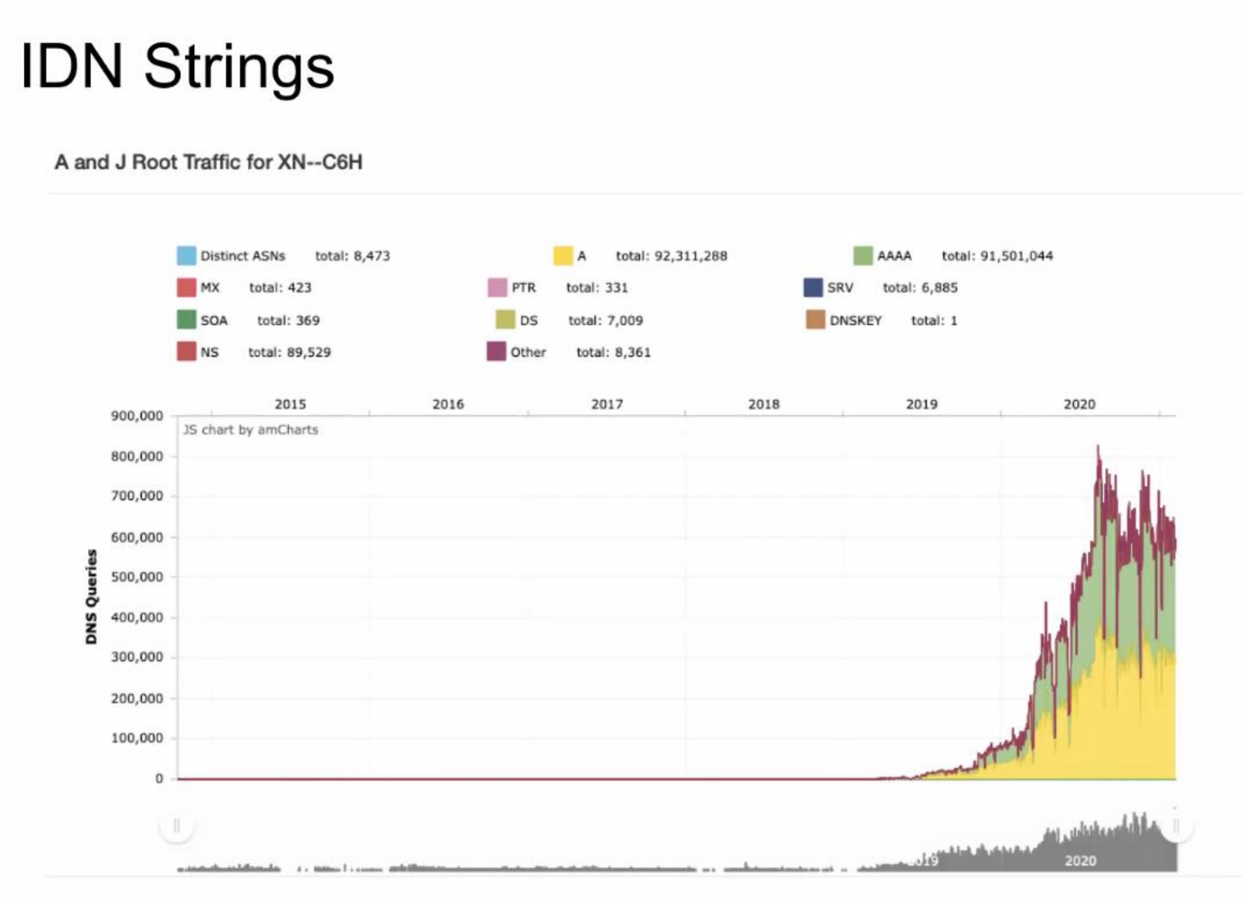
SLIDE 6: String Query Rapid Volume Increase



String Query Rapid Volume Increase

One Top string is j051md46 goes back 2018 then Oct 2020 the traffic becomes 2.5 million quieres per day. Temporal dependency

SLIDE 7: IDN Strings

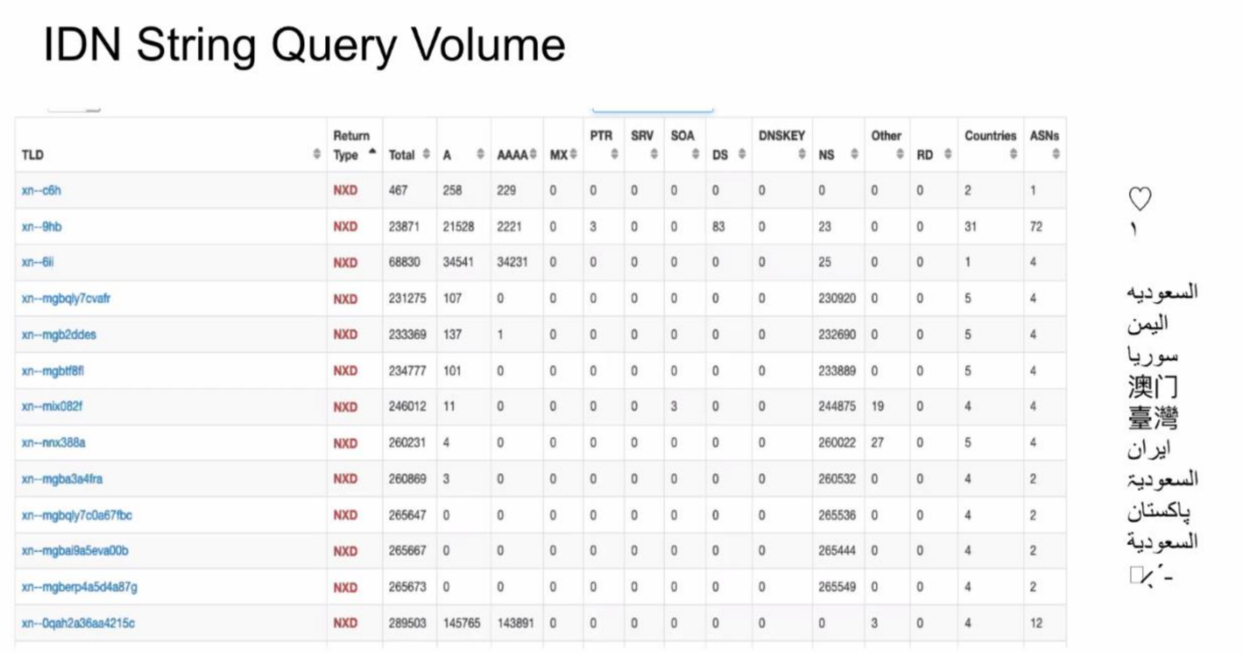


IDN Strings

A and J Root Traffic for XN--C6H

A-Z only search excludes IDNs. Decent amount of IDN strings that are leaking up into the root.

SLIDE 8: IDN String Query Volume



IDN String Query Volume

TLD	Return Type	Total	A	AAAA	MX	PTR	SRV	SOA	DS	DNSKEY	NS	Other	RD	Countries	ASNs
xn--c6h	NXD	467	258	229	0	0	0	0	0	0	0	0	0	2	1
xn--9hb	NXD	23871	21528	2221	0	3	0	0	83	0	23	0	0	31	72
xn--6ii	NXD	68830	34541	34231	0	0	0	0	0	0	25	0	0	1	4
xn--mgbqly7cvafr	NXD	231275	107	0	0	0	0	0	0	0	230920	0	0	5	4
xn--mgb2ddes	NXD	233369	137	1	0	0	0	0	0	0	232690	0	0	5	4
xn--mgbtf8fl	NXD	234777	101	0	0	0	0	0	0	0	233889	0	0	5	4
xn--mix082f	NXD	246012	11	0	0	0	0	3	0	0	244875	19	0	4	4
xn--nnx388a	NXD	260231	4	0	0	0	0	0	0	0	260022	27	0	5	4
xn--mgba3a4fra	NXD	260869	3	0	0	0	0	0	0	0	260532	0	0	4	2
xn--mgbqly7c0a67fbc	NXD	265647	0	0	0	0	0	0	0	0	265536	0	0	4	2
xn--mgbai9a5eva00b	NXD	265667	0	0	0	0	0	0	0	0	265444	0	0	4	2
xn--mgberp4a5d4a87g	NXD	265673	0	0	0	0	0	0	0	0	265549	0	0	4	2
xn--0qah2a36aa4215c	NXD	289503	145765	143891	0	0	0	0	0	0	0	3	0	4	12

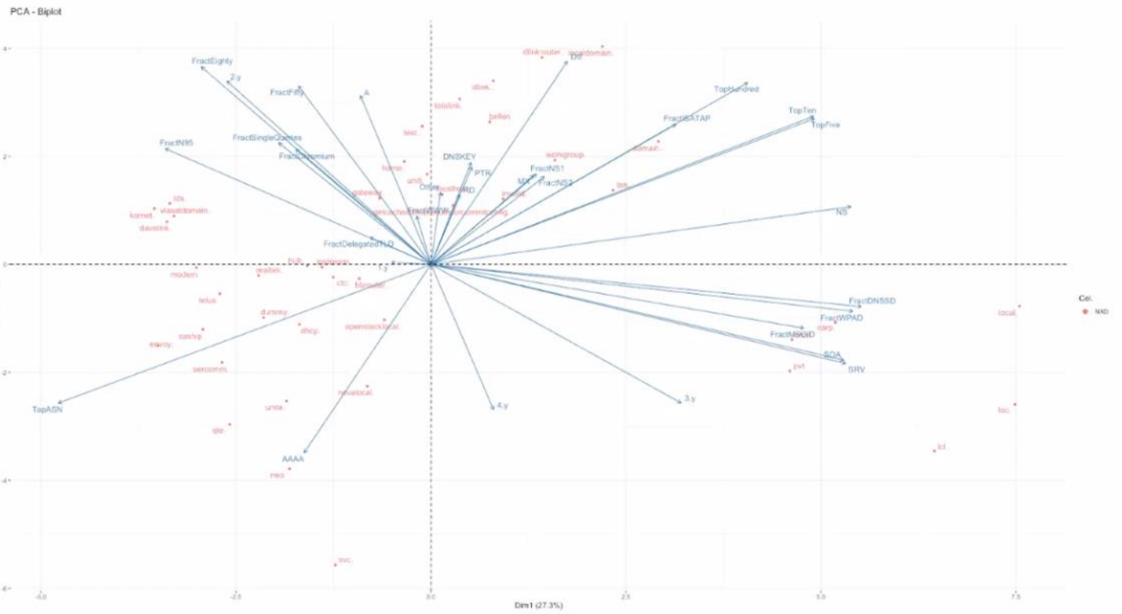
♡
١

السعوديه
اليمن
سوريا
澳门
臺灣
ایران
السعودیۃ
پاکستان
السعودية
□ٰ⁄ ٗ-

xn--c6h = ♡

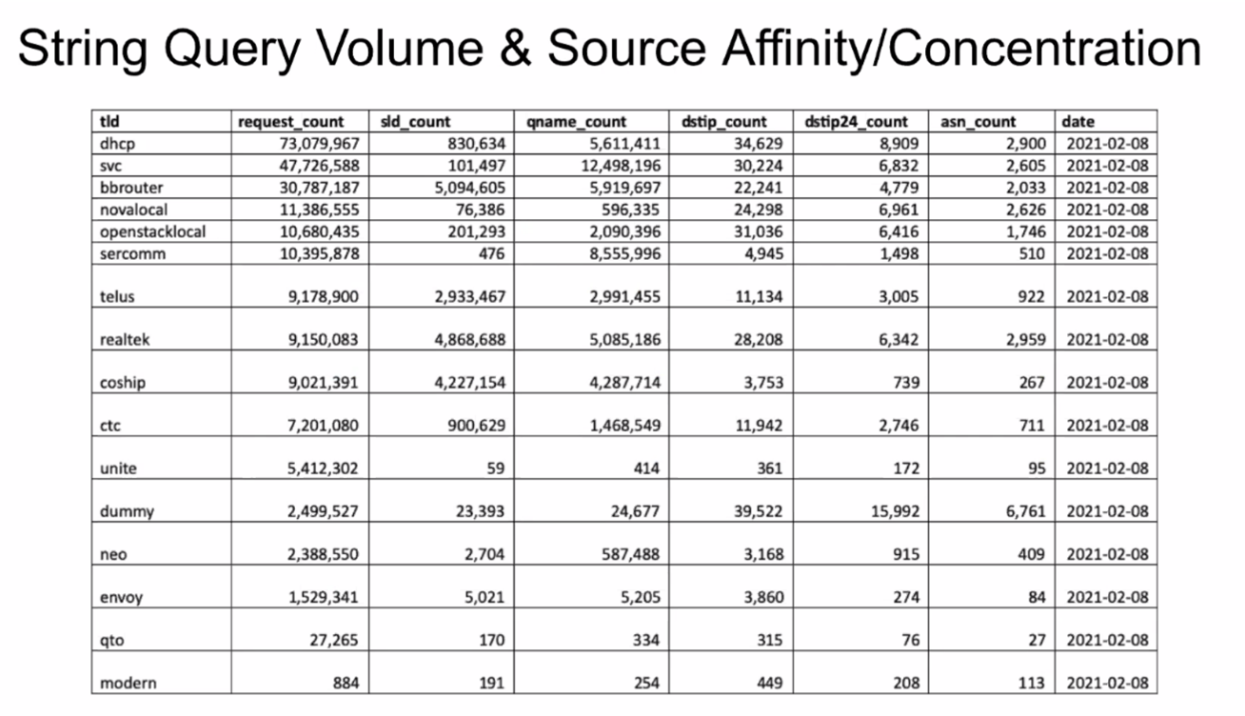
SLIDE 9: PCA Analysis of Top Strings

PCA Analysis of Top Strings



Lower left hand corner: stings focused on traffic heavily anchored or biased within a specific set of ASNs

SLIDE 10: String Query Volume &b Source Affinity/Concentration



String Query Volume & Source Affinity/Concentration

tld	request_count	sld_count	qname_count	dstip_count	dstip24_count	asn_count	date
dhcp	73,079,967	830,634	5,611,411	34,629	8,909	2,900	2021-02-08
svc	47,726,588	101,497	12,498,196	30,224	6,832	2,605	2021-02-08
bbrouter	30,787,187	5,094,605	5,919,697	22,241	4,779	2,033	2021-02-08
novalocal	11,386,555	76,386	596,335	24,298	6,961	2,626	2021-02-08
openstacklocal	10,680,435	201,293	2,090,396	31,036	6,416	1,746	2021-02-08
sercomm	10,395,878	476	8,555,996	4,945	1,498	510	2021-02-08
telus	9,178,900	2,933,467	2,991,455	11,134	3,005	922	2021-02-08
realtek	9,150,083	4,868,688	5,085,186	28,208	6,342	2,959	2021-02-08
coship	9,021,391	4,227,154	4,287,714	3,753	739	267	2021-02-08
ctc	7,201,080	900,629	1,468,549	11,942	2,746	711	2021-02-08
unite	5,412,302	59	414	361	172	95	2021-02-08
dummy	2,499,527	23,393	24,677	39,522	15,992	6,761	2021-02-08
neo	2,388,550	2,704	587,488	3,168	915	409	2021-02-08
envoy	1,529,341	5,021	5,205	3,860	274	84	2021-02-08
qto	27,265	170	334	315	76	27	2021-02-08
modern	884	191	254	449	208	113	2021-02-08

Can a proactive outreach or remediation effort tactually prevent or reduce that risk because you have the ability to talk to a narrower group of perpetrators.

Jeff: Do each of these tlds meet the persistence of 30 days out of 30days?
Matt: yes, assume that.

Jeff: each of these tlds, do you know any of them associated with a particular company, etc?
Matt: some are ISP Manufacturers are easy to identify. Other ones you have to look at ASNs it is leaking from to figure out – find source

DISCUSSION

MEASUREMENTS WE SHOULD CATALOGUE AROUND WHAT QUESTIONS WE SHOULD ASK DATA SOURCES/PROVIDERS?

Jim: now we need to figure out what is next. What other analysis should we do? Review Board questions.

Matt: .corp/.home/.mail were specified for review by the Board.
Rod: create a group work space to start writing down measurements we want and discuss each. If you have a string you see in A & J data or diddle data and there is a chance for minimization do we want to collect a list of those and ask recursive operators about what distribution looks like. About corp/.home/.mail we have a lot of data- lets start testing the measurements we have proposed

Jeff: we are talking about finite set of strings. From that set will se determine what correlations we can make and advise the board for future applications. Do we think there is a need to go down another level to strings that haven't met the high%. Will we decide that below a certain threshold it is safe (or not worth look at?

MATT: qualitative component. What guidance to give when you don't have traffic data to do a risk assessment but the string itself is using a word that has potential risk or harm (like .nuclear)

JIM: unlikely that we can provide a black and white algorithm that the Board can use to judge any case. Instead we will give them tools to use but also need to use their judgement. Regarding Matt's data: