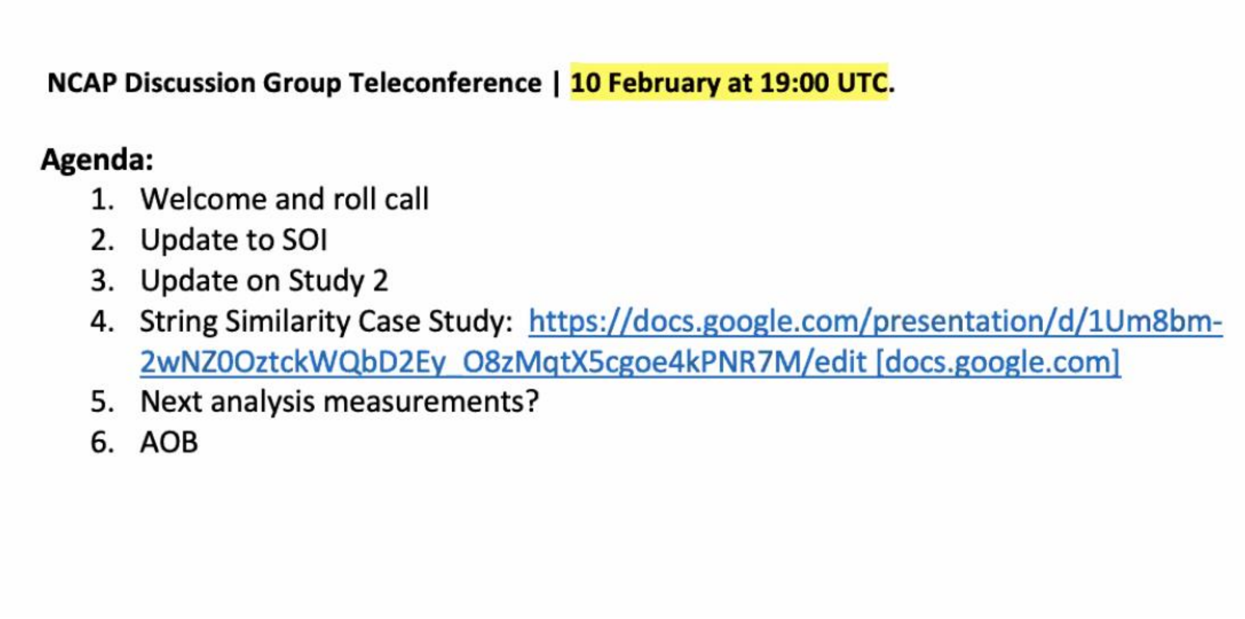


NCAP Discussion Group Teleconference | 10 February at 19:00 UTC.

Agenda:

1. Welcome and roll call
2. Update to SOI
3. Update on Study 2
4. String Similarity Case Study: https://docs.google.com/presentation/d/1Um8bm-2wNZ0OztckWQbD2Ey_O8zMqtX5cgoe4kPNR7M/edit [docs.google.com]
5. Next analysis measurements?
6. AOB

Table of Contents

SOI: 2

Study 2 Update 2

Case Study: Name Collision Analysis Data Exploration 2

Slide 1: Agenda 2

Slide 2: Data Attributes When Evaluating Collision Strings 3

Slide 3: Patterns, Similarity and Clustering 3

Slide 4: K-Means Cluster Overview 4

Slide 5: Collision String Data Features 4

Slide 6: Collision String Similarity/Distance Measurements 5

Slide 7: Collision String Similarity/Distance Measurements 5

Slide 9: Collision String K - Measurements 6

Slide 10: Collision String K-Means 7

Slide 11: Collision String Prin. Comp. Analysis 7

Slide 12: Collision String K-Means – Undelegated Strings 8

Slide 13: Collision String Prin. Comp. Analysis 8

SLIDE 14: Board Questions 9

SOI: no updates

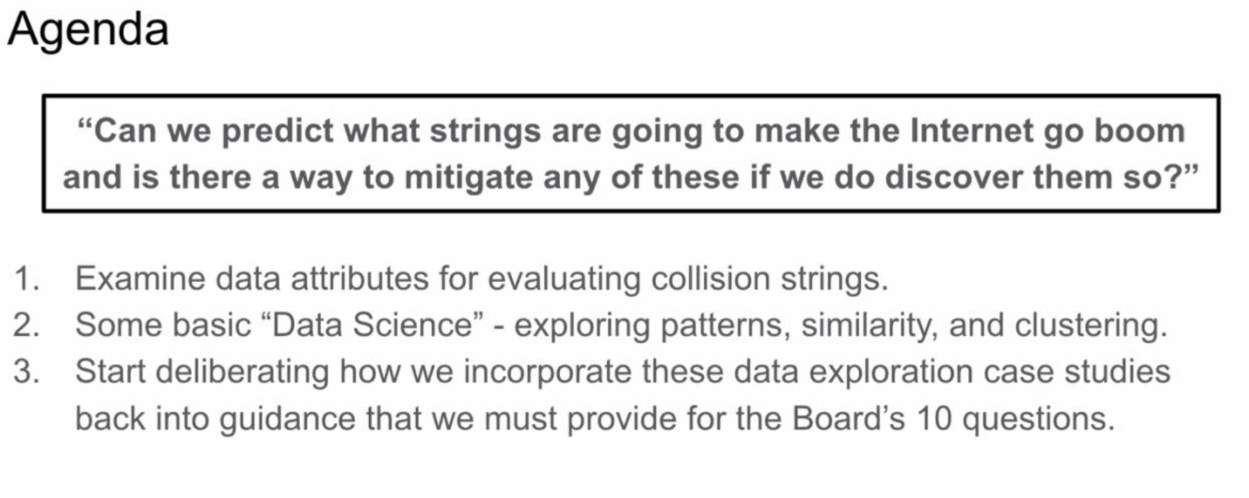
Study 2 Update: still working on final package too send to BTC.

Case Study: Name Collision Analysis Data Exploration

Name Collision Analysis
Data Exploration

took top 80 strings based on query, delegted and undelegated strings

Slide 1: Agenda

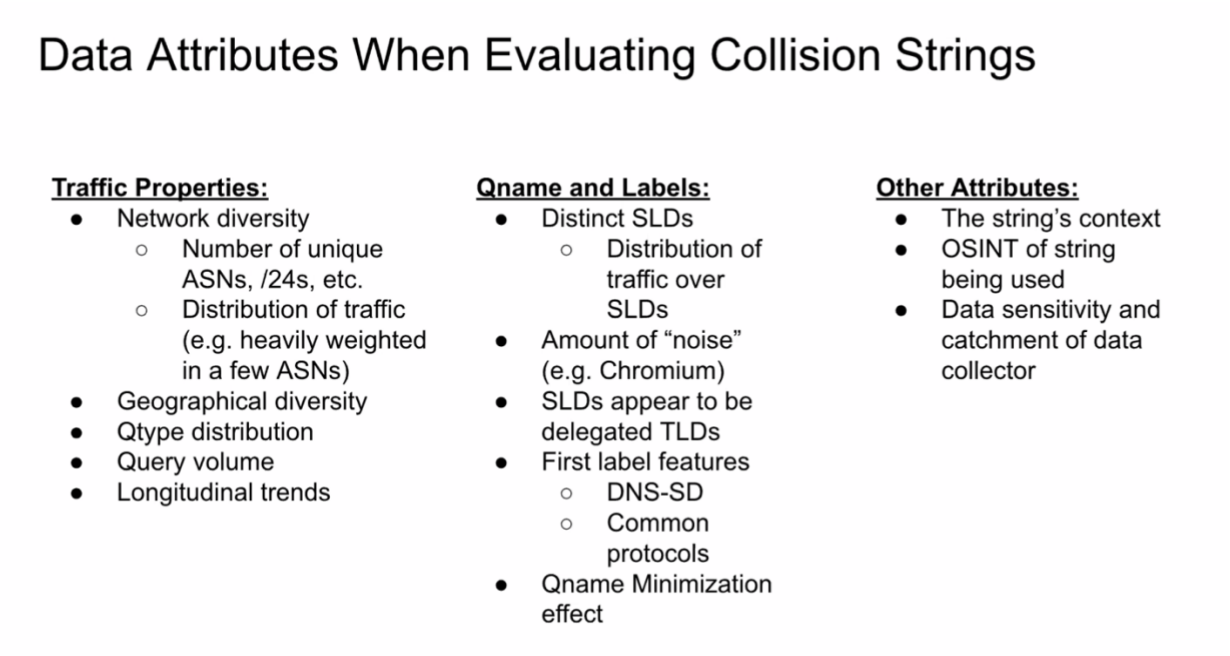


Agenda

> “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. Examine data attributes for evaluating collision strings.
2. Some basic “Data Science” - exploring patterns, similarity, and clustering.
3. Start deliberating how we incorporate these data exploration case studies back into guidance that we must provide for the Board’s 10 questions.

Slide 2: Data Attributes When Evaluating Collision Strings



Data Attributes When Evaluating Collision Strings

Traffic Properties:

- Network diversity
  - Number of unique ASNs, /24s, etc.
  - Distribution of traffic (e.g. heavily weighted in a few ASNs)
- Geographical diversity
- Qtype distribution
- Query volume
- Longitudinal trends

Qname and Labels:

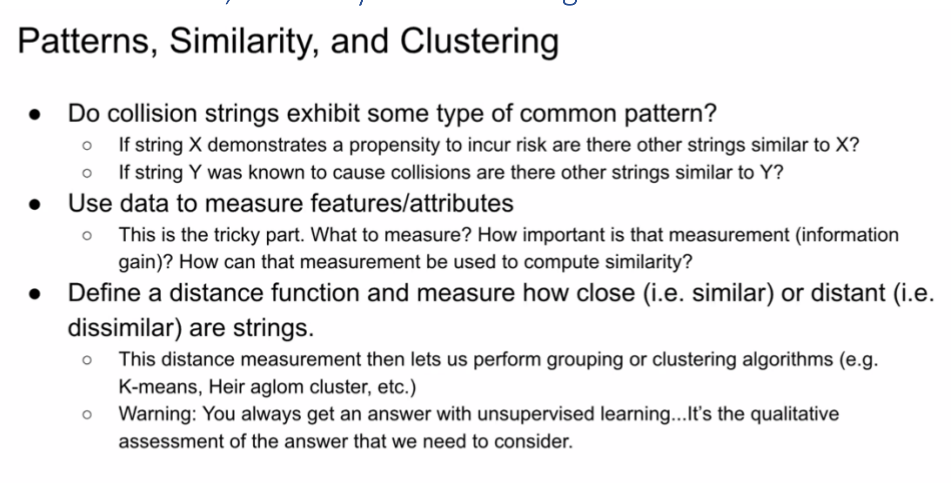
- Distinct SLDs
  - Distribution of traffic over SLDs
- Amount of "noise" (e.g. Chromium)
- SLDs appear to be delegated TLDs
- First label features
  - DNS-SD
  - Common protocols
- Qname Minimization effect

Other Attributes:

- The string's context
- OSINT of string being used
- Data sensitivity and catchment of data collector

Give us Quantitative: Traffic Properties
Gives us context: Qname and labels column
Qualitative attributes is the string's context

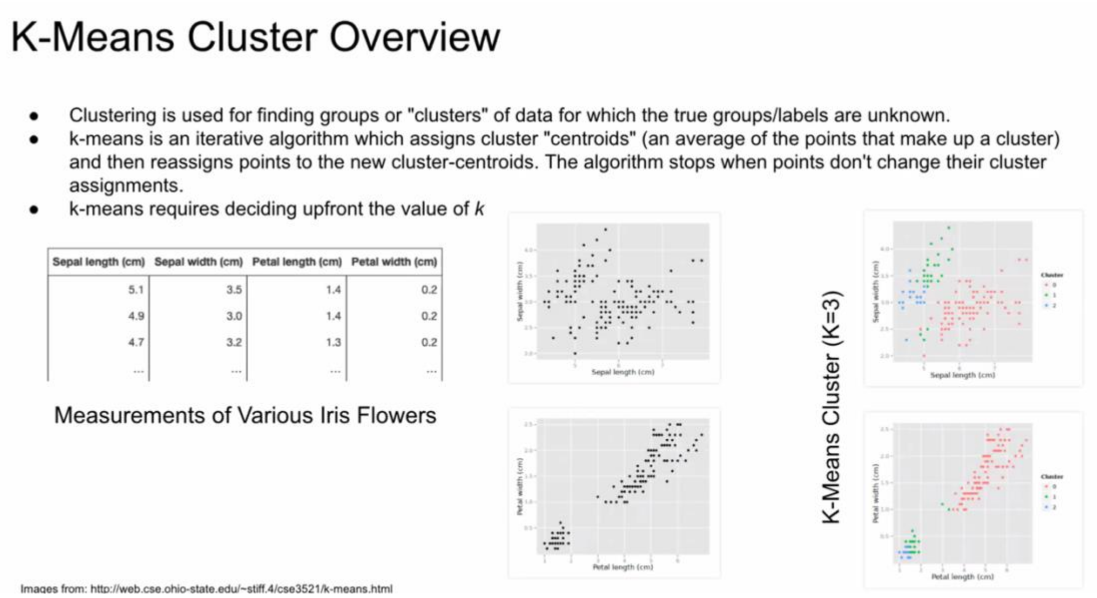
Slide 3: Patterns, Similarity and Clustering



Patterns, Similarity, and Clustering

- Do collision strings exhibit some type of common pattern?
  - If string X demonstrates a propensity to incur risk are there other strings similar to X?
  - If string Y was known to cause collisions are there other strings similar to Y?
- Use data to measure features/attributes
  - This is the tricky part. What to measure? How important is that measurement (information gain)? How can that measurement be used to compute similarity?
- Define a distance function and measure how close (i.e. similar) or distant (i.e. dissimilar) are strings.
  - This distance measurement then lets us perform grouping or clustering algorithms (e.g. K-means, Heir aglom cluster, etc.)
  - Warning: You always get an answer with unsupervised learning...It's the qualitative assessment of the answer that we need to consider.

Slide 4: K-Means Cluster Overview



K-Means Cluster Overview

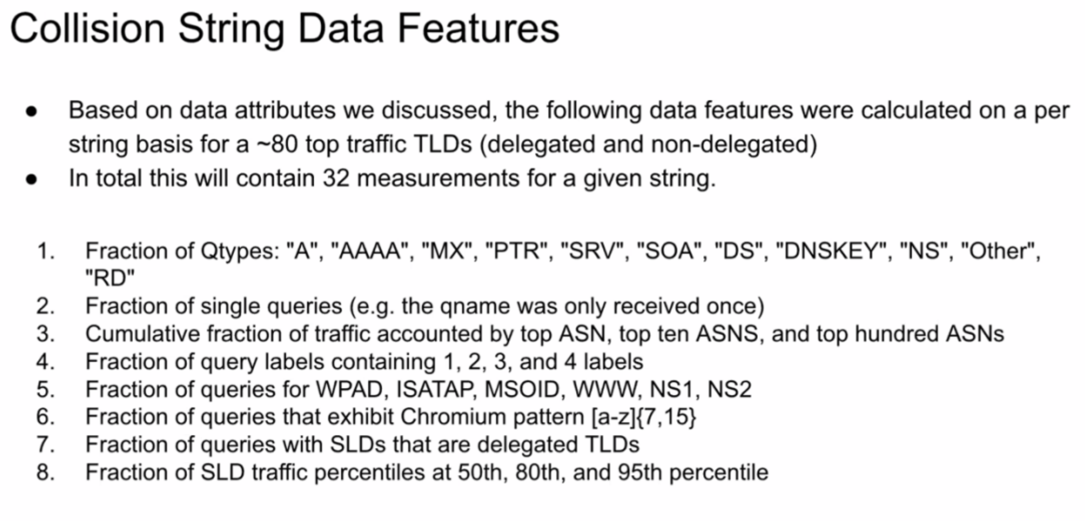
- Clustering is used for finding groups or "clusters" of data for which the true groups/labels are unknown.
- k-means is an iterative algorithm which assigns cluster "centroids" (an average of the points that make up a cluster) and then reassigns points to the new cluster-centroids. The algorithm stops when points don't change their cluster assignments.
- k-means requires deciding upfront the value of *k*

Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
...	...	...	...

Measurements of Various Iris Flowers

Images from: http://web.cse.ohio-state.edu/~stiff.4/cse3521/k-means.html

Slide 5: Collision String Data Features

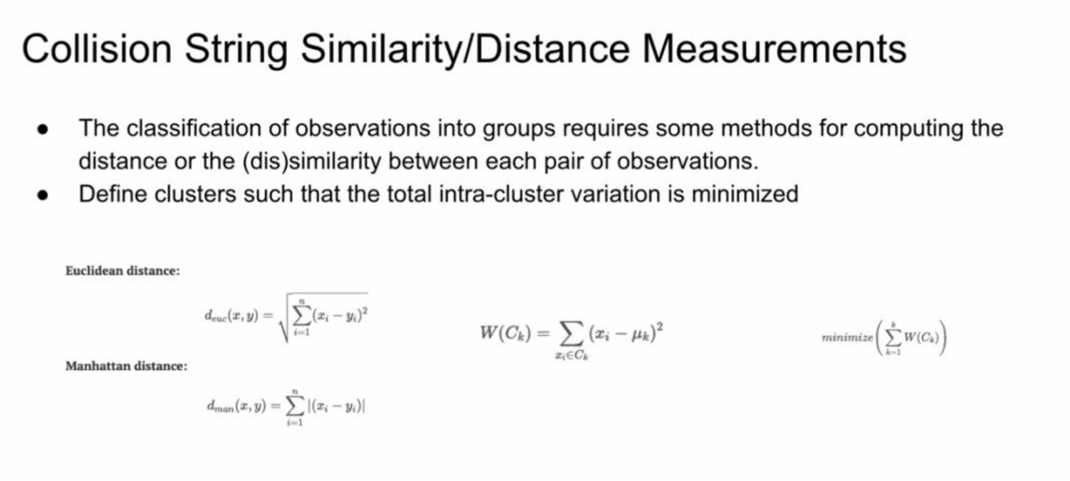


Collision String Data Features

- Based on data attributes we discussed, the following data features were calculated on a per string basis for a ~80 top traffic TLDs (delegated and non-delegated)
- In total this will contain 32 measurements for a given string.

1. Fraction of Qtypes: "A", "AAAA", "MX", "PTR", "SRV", "SOA", "DS", "DNSKEY", "NS", "Other", "RD"
2. Fraction of single queries (e.g. the qname was only received once)
3. Cumulative fraction of traffic accounted by top ASN, top ten ASNS, and top hundred ASNs
4. Fraction of query labels containing 1, 2, 3, and 4 labels
5. Fraction of queries for WPAD, ISATAP, MSOID, WWW, NS1, NS2
6. Fraction of queries that exhibit Chromium pattern [a-z]{7,15}
7. Fraction of queries with SLDs that are delegated TLDs
8. Fraction of SLD traffic percentiles at 50th, 80th, and 95th percentile

Slide 6: Collision String Similarity/Distance Measurements



Collision String Similarity/Distance Measurements

- The classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations.
- Define clusters such that the total intra-cluster variation is minimized

Euclidean distance:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

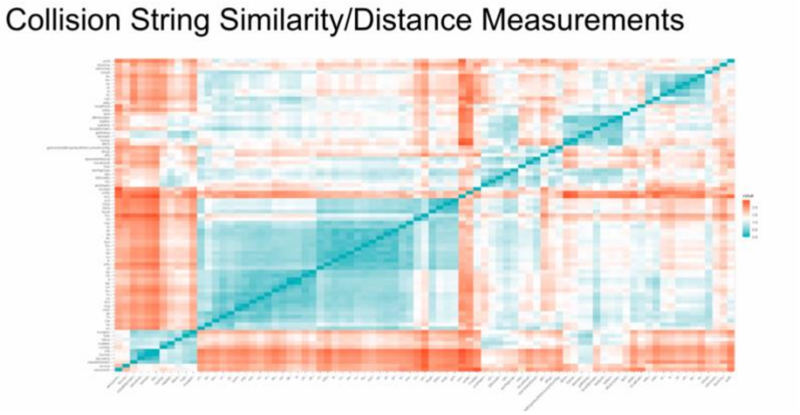
$$W(C_k) = \sum_{x_i \in C_k}(x_i - \mu_k)^2$$

$$minimize\left(\sum_{k=1}^{k}W(C_k)\right)$$

Manhattan distance:

$$d_{man}(x,y) = \sum_{i=1}^{n}|(x_i - y_i)|$$

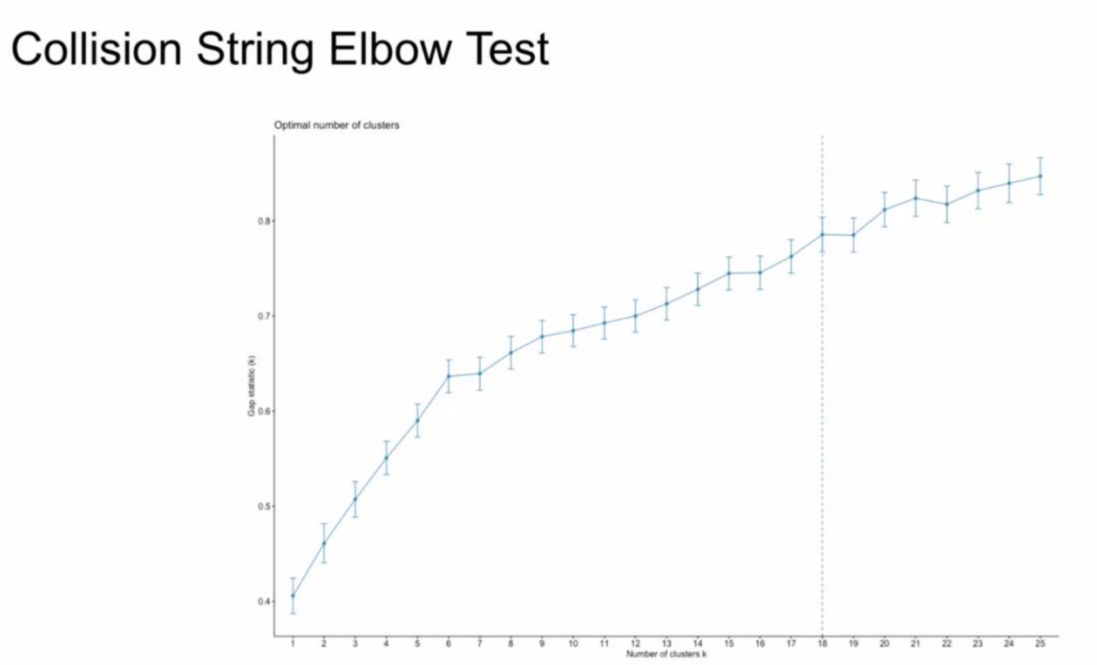
Slide 7: Collision String Similarity/Distance Measurements



Collision String Similarity/Distance Measurements

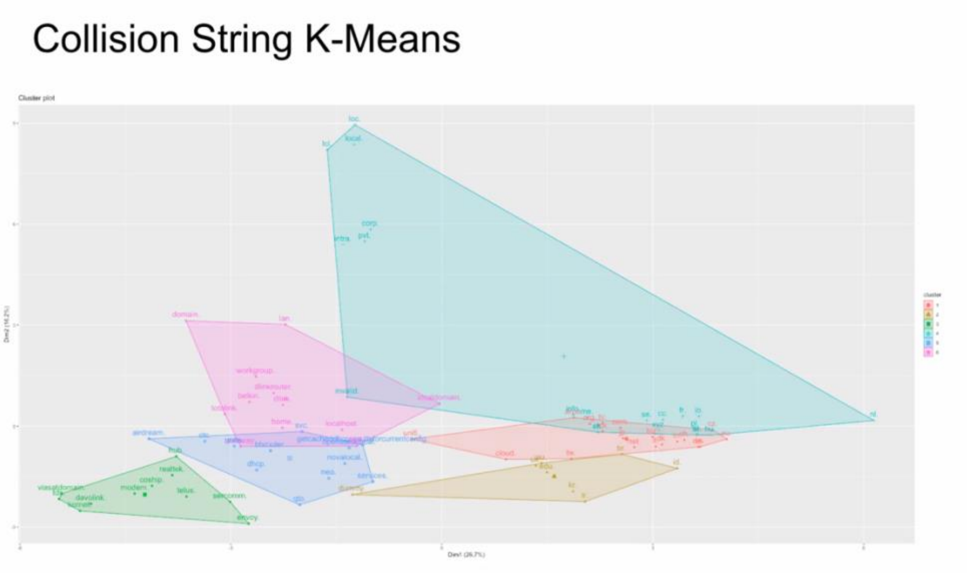
Each square is how similar are 2 strings based on 32 attributes. Blue squares as patterns emerge, red boxes are saying these strings are very far apart.

Slide 8: Collision String Elbow Test


Collision String Elbow Test

How do you pick a K? This is exercise of computing all the # of different Ks up to 25 and trying to find the ‘elbow’ in the curve.

Slide 9: Collision String K - Measurements

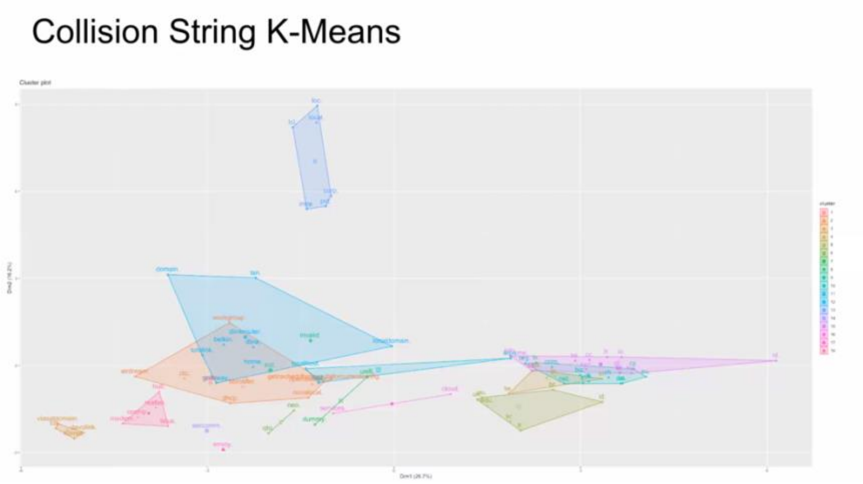

Collision String K-Means

Result of clustering. K=6, so 6 clusters.
X and y axis are called dimension 1 and 2.
Each string has 32 data attributes.
To visualize K use principal component analysis that looks for ways of reducing the variance and transposing that and projecting that onto a dimensional graph
X is accounting for 26.7% of variance
Y is accounting for 16.2% of variance

Both together = 43% of total variance, so they are primary components

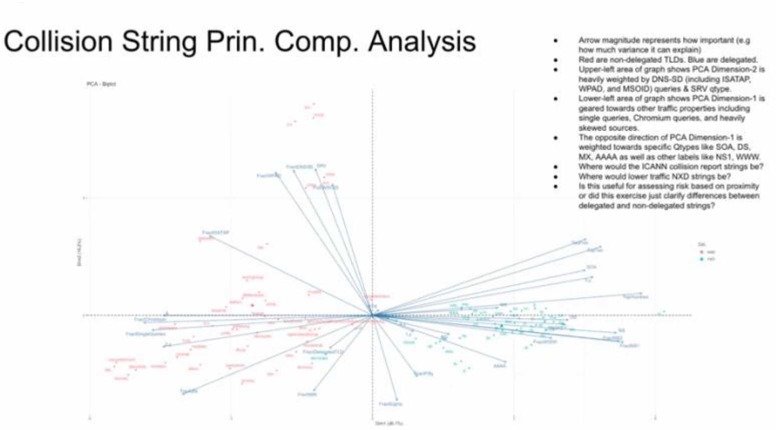
Group 1 in lower left hand corner you have real tech and telis all clustering together they are all home residential ISP routers.
Group 2: Next to group 1 it is blue cluster with similar strings grouped together
Group 3: Pink cluster is local LANS
Group 4: Right side is all delegated TLDS – ccTLD and GTLDS are together

Slide 10: Collision String K-Means



Collision String K-Means

Slide 11: Collision String Prin. Comp. Analysis

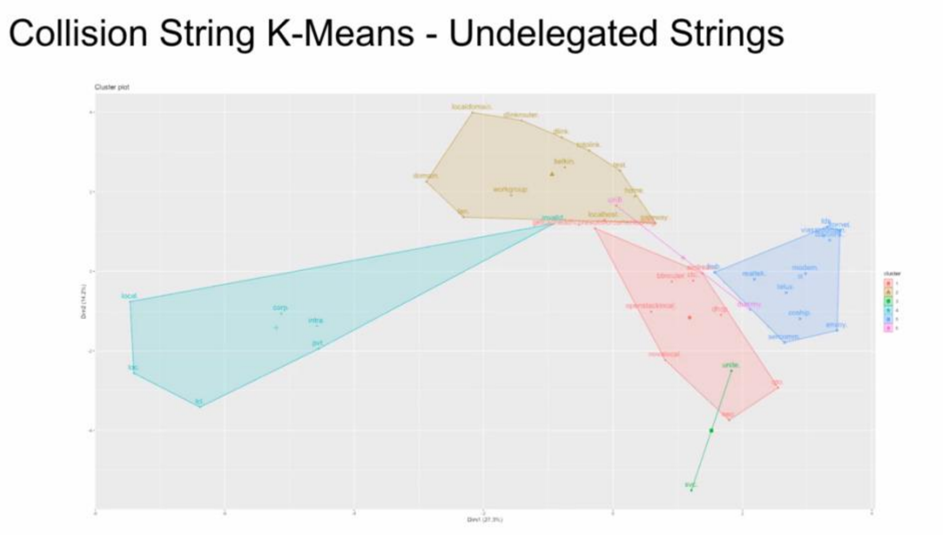


Collision String Prin. Comp. Analysis

- Arrow magnitude represents how important (e.g how much variance it can explain)
- Red are non-delegated TLDs. Blue are delegated.
- Upper-left area of graph shows PCA Dimension-2 is heavily weighted by DNS-SD (including ISATAP, WPAD, and MSOID) queries & SRV qtype.
- Lower-left area of graph shows PCA Dimension-1 is geared towards other traffic properties including single queries, Chromium queries, and heavily skewed sources.
- The opposite direction of PCA Dimension-1 is weighted towards specific Qtypes like SOA, DS, MX, AAAA as well as other labels like NS1, WWW.
- Where would the ICANN collision report strings be?
- Where would lower traffic NXD strings be?
- Is this useful for assessing risk based on proximity or did this exercise just clarify differences between delegated and non-delegated strings?

31 mins in:
Upper left hand corner

Slide 12: Collision String K-Means – Undelegated Strings

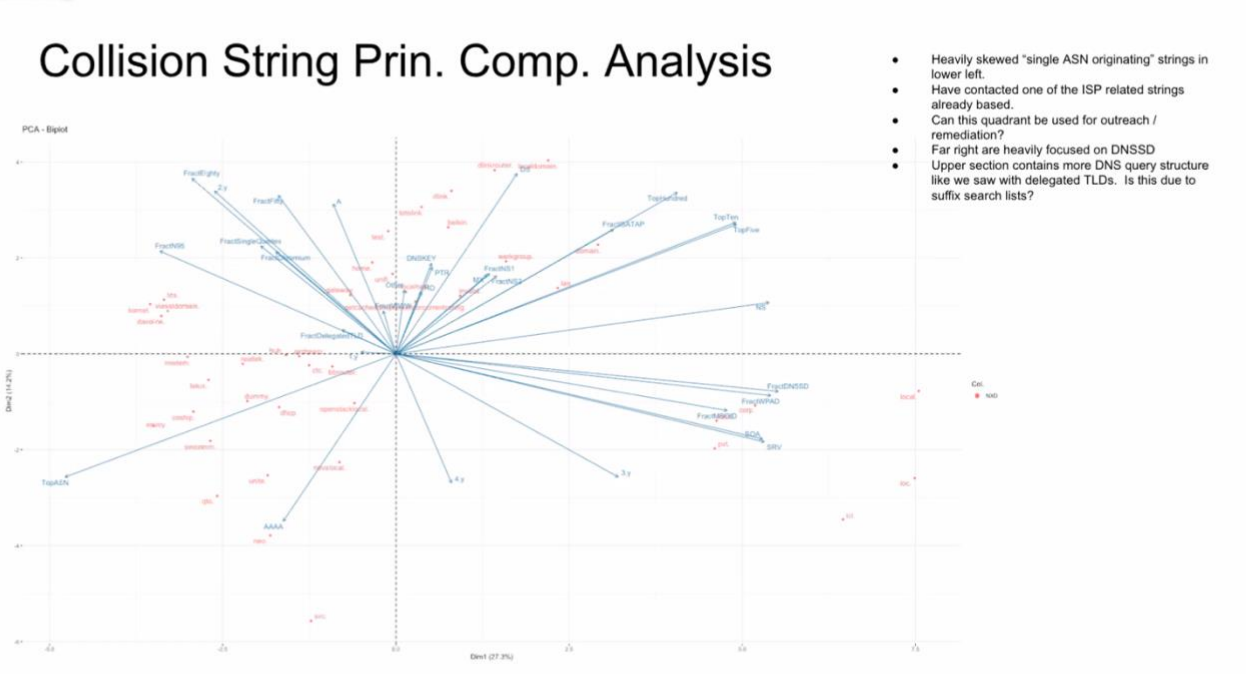


Collision String K-Means - Undelegated Strings

K=6

Home routers grouping together

Slide 13: Collision String Prin. Comp. Analysis



Collision String Prin. Comp. Analysis

- Heavily skewed "single ASN originating" strings in lower left.
- Have contacted one of the ISP related strings already based.
- Can this quadrant be used for outreach / remediation?
- Far right are heavily focused on DNSSD
- Upper section contains more DNS query structure like we saw with delegated TLDs. Is this due to suffix search lists?

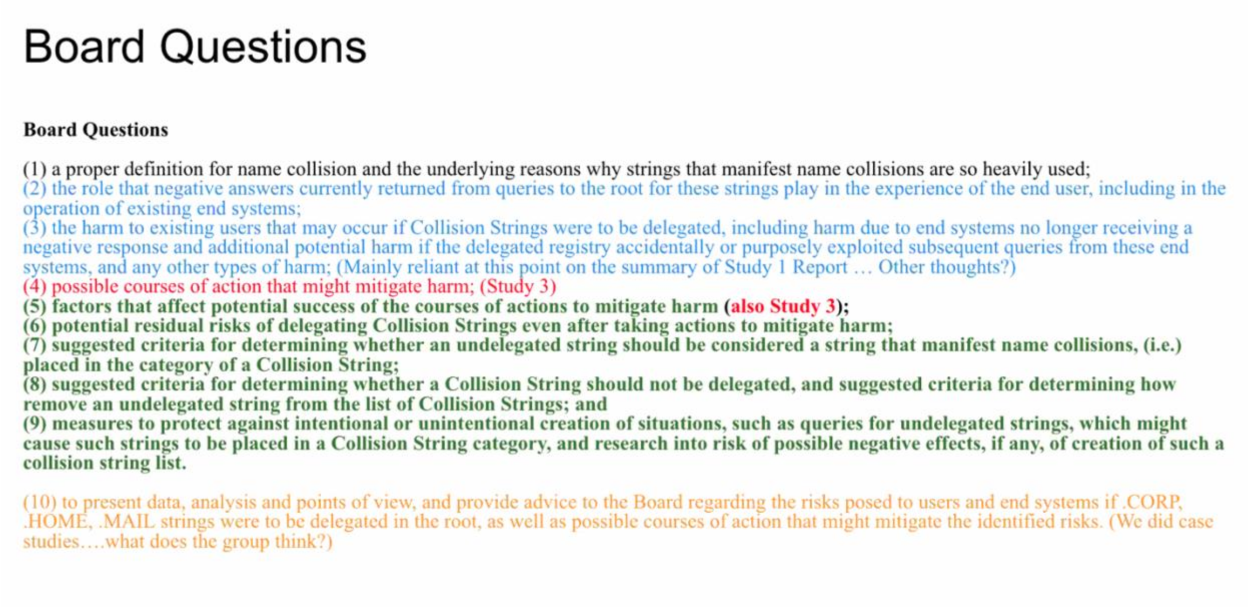
Lower left was influenced by top ASN – 1 source.

Residential ISP Matt has reached out to and knows they are leaking queries, but it is front-ware updates since it's home routers and is a long process to fix.

Is there a way to identify ones that are more likely to be able to fix, vs. upper right hand corner where traffic is spread out from many sources and would be hard to remediate.

Would those strings that are delegated have looked like the undelegated strings prior to their delegation?

SLIDE 14: Board Questions



Board Questions

Board Questions

(1) a proper definition for name collision and the underlying reasons why strings that manifest name collisions are so heavily used;
(2) the role that negative answers currently returned from queries to the root for these strings play in the experience of the end user, including in the operation of existing end systems;
(3) the harm to existing users that may occur if Collision Strings were to be delegated, including harm due to end systems no longer receiving a negative response and additional potential harm if the delegated registry accidentally or purposely exploited subsequent queries from these end systems, and any other types of harm; (Mainly reliant at this point on the summary of Study 1 Report … Other thoughts?)
(4) possible courses of action that might mitigate harm; (Study 3)
(5) factors that affect potential success of the courses of actions to mitigate harm (also Study 3);
(6) potential residual risks of delegating Collision Strings even after taking actions to mitigate harm;
(7) suggested criteria for determining whether an undelegated string should be considered a string that manifest name collisions, (i.e.) placed in the category of a Collision String;
(8) suggested criteria for determining whether a Collision String should not be delegated, and suggested criteria for determining how remove an undelegated string from the list of Collision Strings; and
(9) measures to protect against intentional or unintentional creation of situations, such as queries for undelegated strings, which might cause such strings to be placed in a Collision String category, and research into risk of possible negative effects, if any, of creation of such a collision string list.

(10) to present data, analysis and points of view, and provide advice to the Board regarding the risks posed to users and end systems if .CORP, .HOME, .MAIL strings were to be delegated in the root, as well as possible courses of action that might mitigate the identified risks. (We did case studies….what does the group think?)

#1 is answered
#2 & 3: from Study 1

How do we take our analysis and create a list of 10 things we want data to answer, and then reach out to other data sources to get more information

A proper definition for name collision: time to reconsider it to updated based on new discoveries?