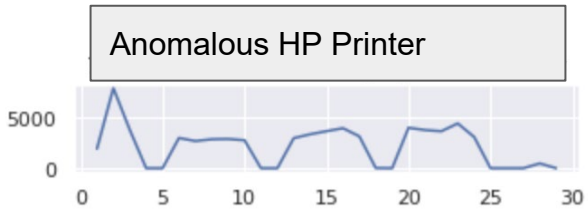
Infoblox

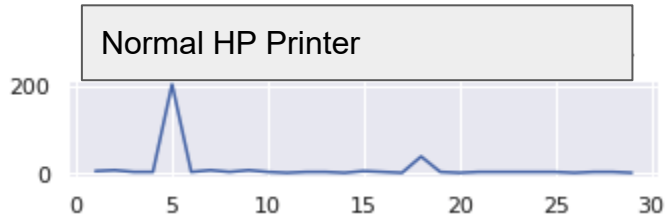
Anomaly Detection on Devices using DNS Queries

Prepared by: Fatemeh Riahi

Problem Definition

- IOT/ICS devices are more likely to get compromised.
  - According to a research by Forrester, 67% of enterprises have experienced IoT security incidents.
- Detect anomalous behaviour of IoT devices.
  - If we track DNS activities of devices and detect anomalous behaviours, we can quickly quarantine such devices.





Solution

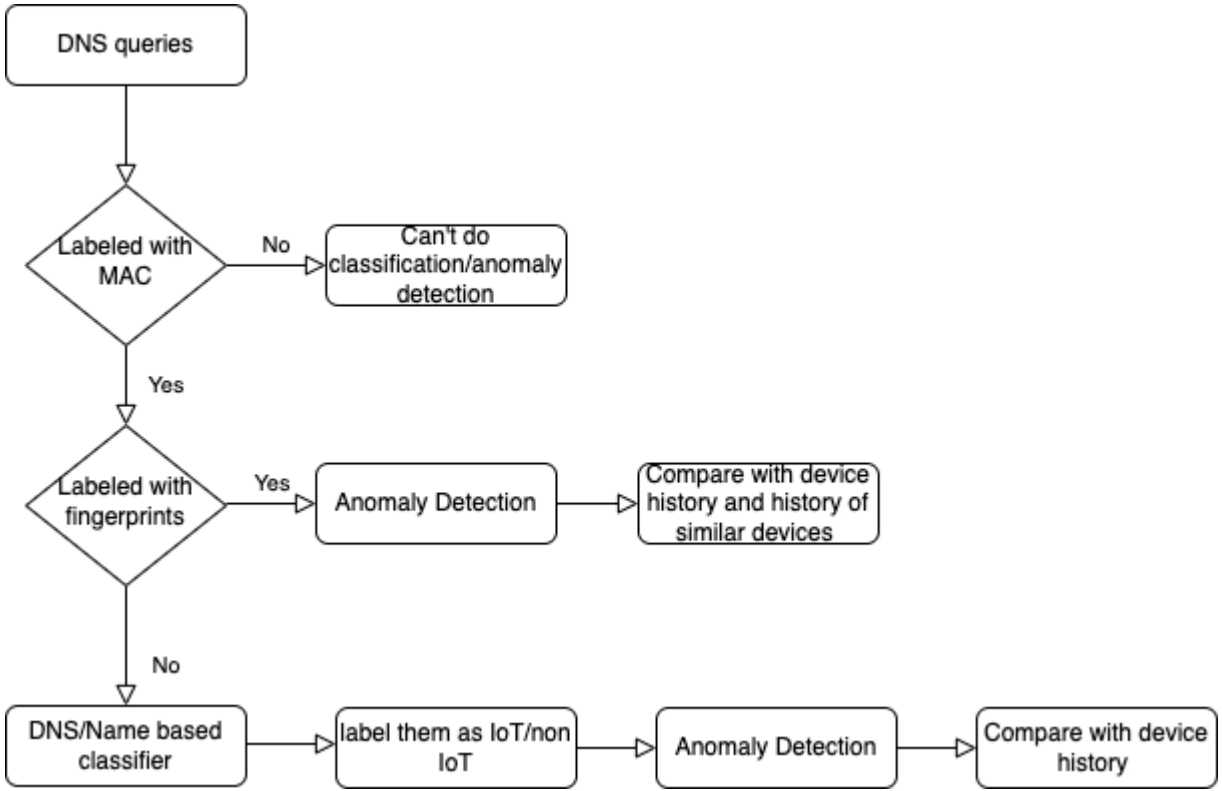
- Find devices that are autonomous using available information.
  - DNS Queries
  - Device Name
  - Device MAC address
- Make a profile for each device that belong to IoT category.
- Vectorize the DNS activities using NLP methods.
- Compare DNS activities of a device with its history and the history of devices with similar fingerprints if the fingerprint is available and detect anomalies.

DNS and DHCP

Field Name	Dynamic/Fixed	Source	Example
IP Address	Dynamic	DNS/DHCP	192.168.17.23
Device Name	Fixed	DHCP	Vinods-Macbook
MAC Address	Fixed/Dynamic	DHCP	aa:bb:cc:11:22:33
Fingerprint*	Fixed	DHCP	MacOS
qname	NA	DNS	netflix.com

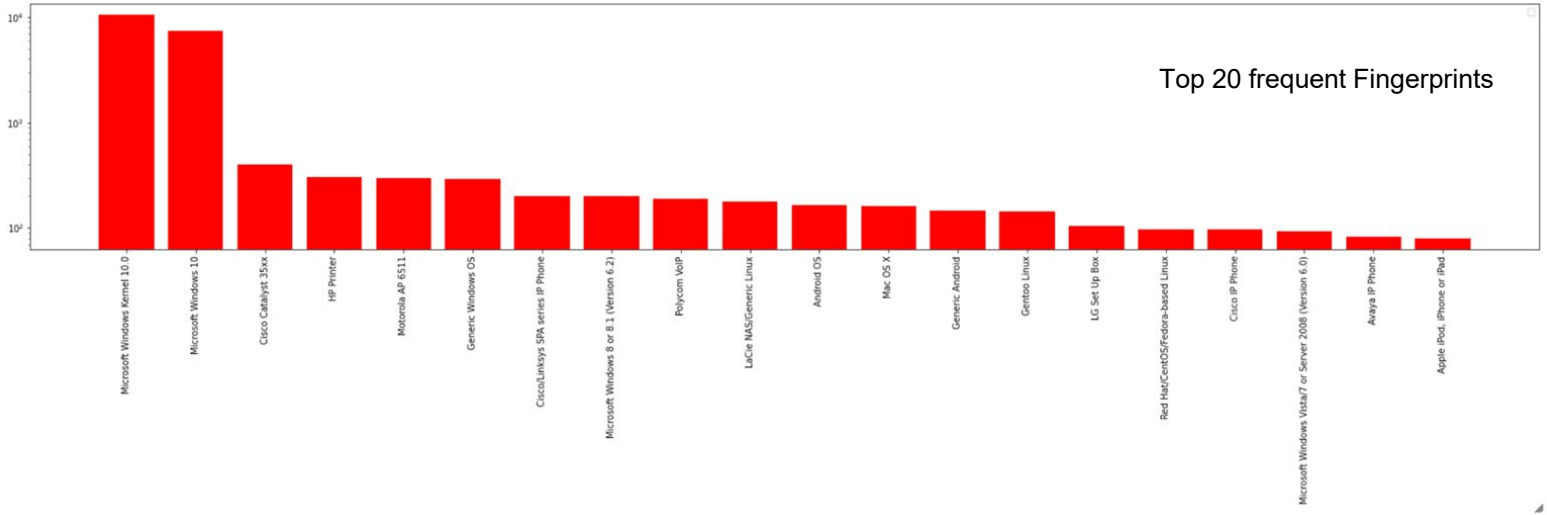
* Fingerprint is a device identifier that is assigned to a device by DHCP server. It can be as general as HP device or as specific as Microsoft Windows 10.

Solution Overview



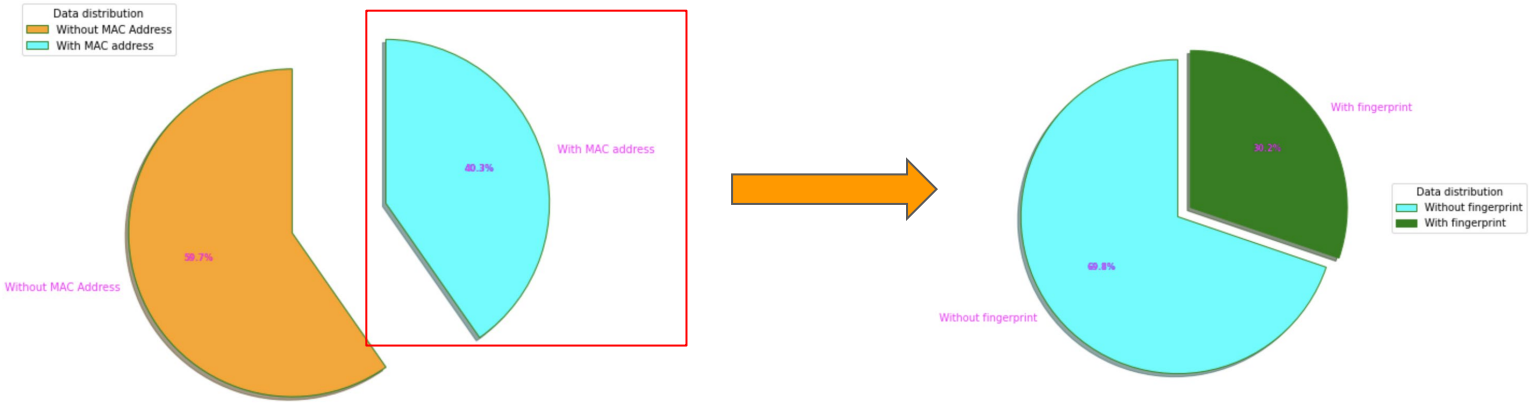
EDAs

- Two billion DNS queries recorded daily.
- Device Identifier cannot be IP address.
  - Using MAC address as device identifier.
- 1.2 million devices in total.
- Some devices are being fingerprinted by DHCP service. But not all of them.

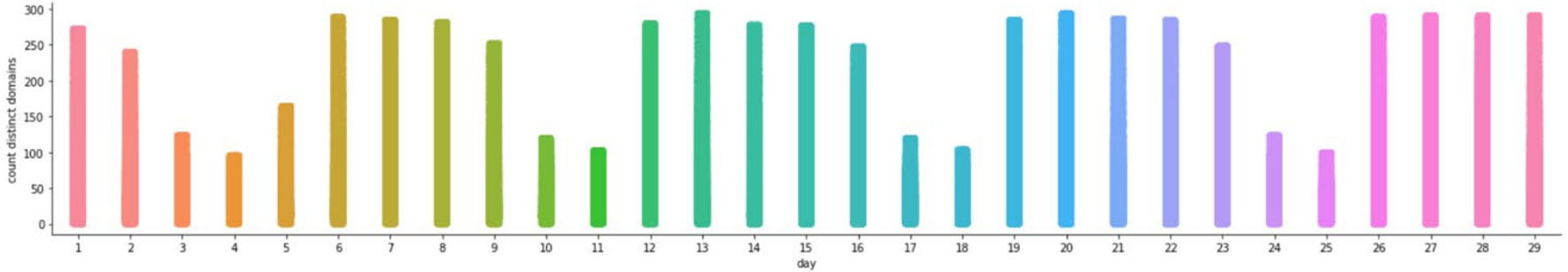


Data Limitation

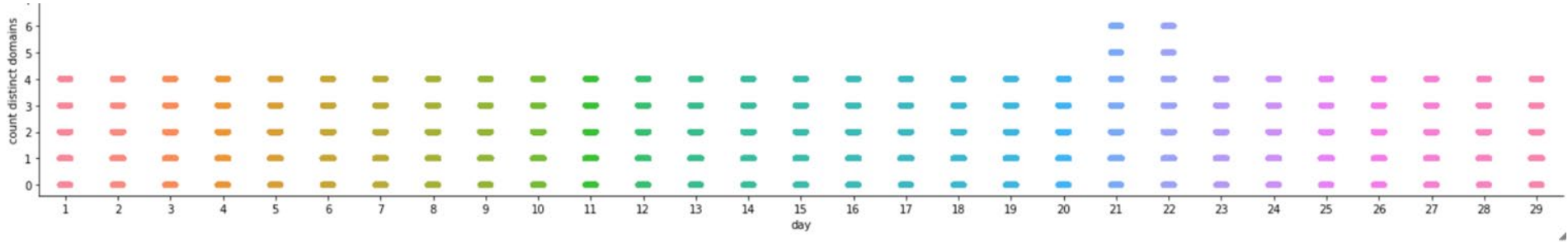
- Some of the DNS queries are tagged by MAC address and Fingerprints but not all.
  - If MAC addresses are not available we don’t have device identifier and we cannot do classification and anomaly detection.
  - For unknown devices we have to do a classification step first.



IoT/ non IoT devices query pattern in a month



One month of activity of non iot devices



One month of activity of iot devices

Feature Construction and Vectorization

Vectorizing and Word Embedding

- Word embedding is a vectorize representation of texts.
  - Each word will be transformed into numerical representation.
    - TFIDF
    - CountVectorizer
    - Word2Vec
- Convert daily DNS queries of each device to a vector.

TFIDF and Countvectorizer

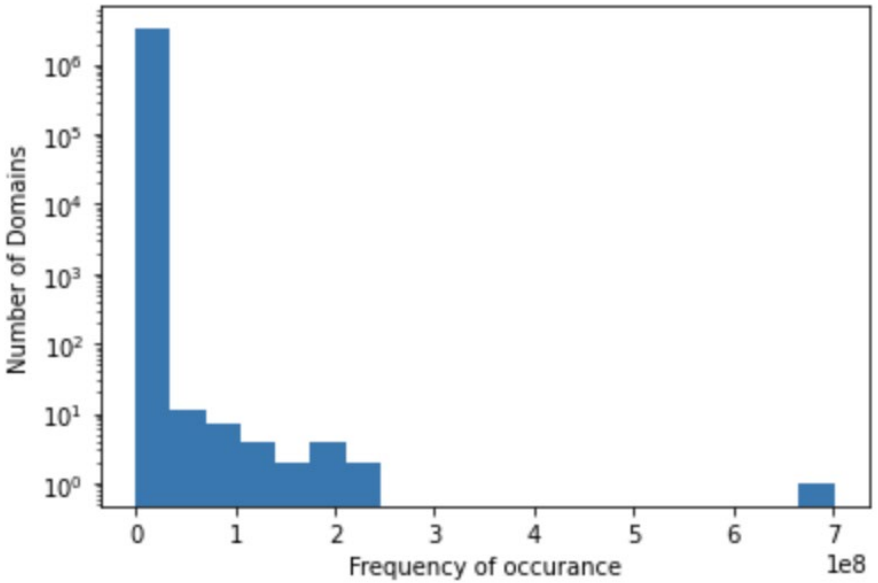
- Term Frequency-Inverse Document Frequency.
  - Inverse document frequency is specially important here because we want to lower the effect of domains like google.com that are queried by most devices.

	mac_address	domains_visited	fingerprint	count_vector	tfidf
0	aa:bb:cc:11:22:ff	google.com facebook.com google.com netflix.com	MS Windows	[0, 1, 2, 0, 1]	[0.0, 0.39, 0.77, 0.0, 0.51]
1	gg:aa:aa:11:22:44	hp.com	HP Printer	[0, 0, 0, 1, 0]	[0.0, 0.0, 0.0, 1.0, 0.0]
2	ss:rr:33:22:11:er	google.com facebook.com apple.com	Macbook	[1, 1, 1, 0, 0]	[0.68, 0.52, 0.52, 0.0, 0.0]

- Countvectorizer: simply counts the frequency of each domain in each device.

Problem of Vectorizers in our Data

- Sparse vector space.
  - Three million unique domains.
  - Small percentage of domains occur most of the time.
  - Too many domains occur only few times .



Word2Vec: How do you define word embedding?

- Word2vec is one of the most common methods of generating word embedding.
- We will define a set of features for describing the words and for each word it will learn the value to each of those features.

King	Queen	Man	Woman	Horse
Authority=1 Gender= -1 Has_tail=0	Authority=1 Gender= 1 Has_tail=0	Authority=0.5 Gender= -1 Has_tail=0	Authority=0.5 Gender= 1 Has_tail=0	Authority=0 Gender= 0 Has_tail=1

King - Man + Woman= Queen
[1,-1,0]-[0.5,-1,0]+[0.5,1,0]= [1,1,0]

Word2Vec

- Sentence-> list of sorted dns queries from one device in a day.
- Word->domain
- corpus->all unique domain
- document -> collection of all the sentences

```
1
2  model.wv.most_similar('wholefoodsmarket.com
```

```
Out[28]: [('acx.com', 0.9595616459846497),
 ('ring.com', 0.9560766816139221),
 ('comixology.com', 0.9559093117713928),
 ('eero.com', 0.9522551894187927),
 ('carbontrust.com', 0.9482857584953308),
 ('boxofficemojo.com', 0.9432862997055054),
 ('blinkforhome.com', 0.9390537142753601),
 ('fabric.com', 0.937058687210083),
 ('fountain.com', 0.9350566864013672),
```

Device Classification

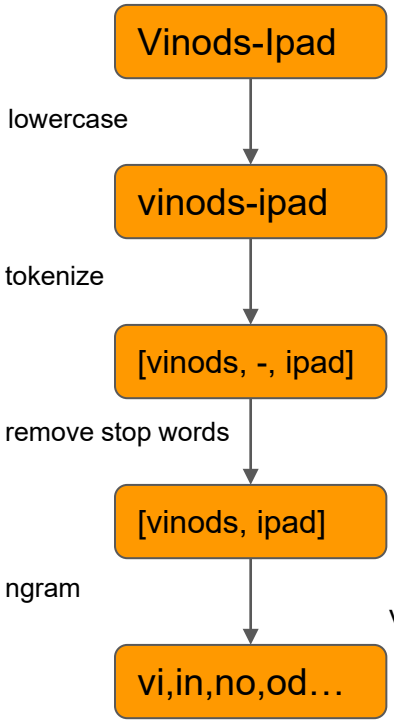
Classifier

- We need a classifier to label the unknown devices for us.
  - Vectorized DNS Queries
  - Use Device Name

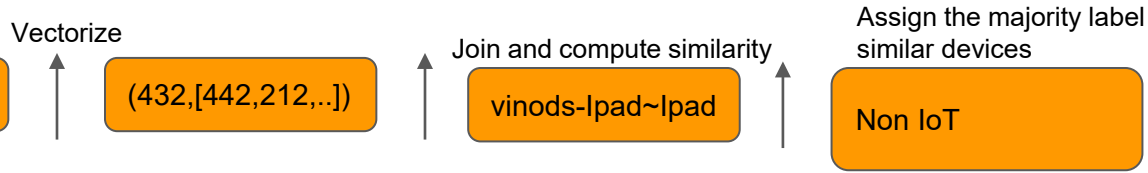
DNS based classifier

- We combine different vectorizers with different well known classifiers.
  - Best performing classifier: Histogram Gradient Boosting.
- TFIIDF vocab size=10000
- While building the vectors with TFIDF only took few seconds, the word2vec took 25 minutes to build the vector for our data.

Name based Classifier



left_mac_address	left_dhcp_fingerprint	left_device_name	similar_dhcp_fp	similar_device_name
00:60:01:81:f1:c8	nomatch	Vinods-iPad	{'Apple OS', 'nomatch'}	{'Vinods-iPad', 'iPad', 'Vinods-iPad-2'}
00:60:01:82:11:e9	nomatch	Vinods-iPad	{'Apple OS', 'nomatch'}	{'Vinods-iPad', 'iPad', 'Vinods-iPad-2'}
28:77:f1:f2:37:f7	nomatch	iPaddeSIFQAArea	{'Apple OS', 'nomatch'}	{'iPaddeSIFQAArea', 'iPaddeSTestArea'}
55:4e:00:01:00:00	nomatch	ONW-CSUNG-01	{'Microsoft Windows Kernel 10.0', 'nomatch'}	{'ONW-CSUNG-01'}
55:4e:00:01:00:00	nomatch	ONW-NPETRESC-01	{'nomatch', 'Fedora 14 based distro'}	{'onx-npetresc-01', 'ONW-NPETRESC-01'}



Different classifier performance

- Recall is more important for our use case(anomaly detection)
- We can only use the Name-based classifier for small portion of our data(28%).

Method	Precision	Recall	Caveats
Name-based Classifier	1	0.99	Only 28% of devices have names
DNS-based with TFIDF	0.92	0.83	Quick to train, we have to limit the vector size to fit the memory and that means majority of domains will be treated as out of bags
DNS based with Word2Vec	0.81	0.92	Slow to train.

Anomaly Detection

Anomaly detection

When an event is considered as anomaly?

- Compare the DNS queries of a device to all the similarly fingerprinted devices on the same customer
- Compare a devices traffic with itself over a period of time

Anomaly Models:

- Create TFIDF, Word2Vec vectors
  - Apply a one class SVM and Isolation forest.
  - Compute cosine similarities between different dates and anything above three standard deviation away from the mean is considered anomalies.

Validation of Anomaly Detectors

- We don’t have labels to show us actual anomalous events in the past.
- The only way for us to validate the models is to evaluate the anomalous events and see if they make sense.
- Low false positive is critical for us here.

Anomaly Results

- Any increase in frequency of known domains was returned as anomalies.
- We take one month of DNS queries for five fingerprints.
  - "Polycom Conference IP Phone","Avaya IP Phone","HP Printer","Cisco IP Phone","Lexmark Printer"
  - 1652 events returned in total, too many!

	mac_address	domains_queried	
0	aa:bb:cc:11:22:33	'{lexmark.com:200,time.com:300}'	✗
1	dd:ee:ff:44:55:66	'{airbnb.com:6,jerrysfoods.com}'	✓

- Stop word removal with regard of each known fingerprints.
- 41 anomalous events that they all looked anomalous and needed investigation.

Example 1 – Anomalous events detected for certain HP printers

- Median number of unique domains

  Anomaly device – 976

  HP Printer - 3

- Median number of unique domains

  Anomaly device – 115

  HP Printer - 3

Anomalous HP printer

'1rx.io', '2mdn.net', '33across.com', '360yield.com', '3lift.com', '4dex.io', 'None', 'a-mo.net', 'a-mx.com', 'a3cloud.net', 'acuityplatform.com', 'ad-m.asia', 'ad-stir.com', 'addthis.com', 'adentifi.com', 'adform.net', 'adgrx.com', 'adingo.jp', 'adition.com', 'adkernel.com', 'admanmedia.com', 'admedo.com', 'admixer.net', 'adnxs-simple.com', 'adnxs.com', 'adobe.com', 'adobe.io', 'adotmob.com', 'adpone.com', 'adrta.com', 'adsafeprotected.com', 'adscience.nl', 'adsrvr.org', 'adsymptotic.com', 'adtdp.com', 'advangelists.com', 'adventori.com', 'advertising.com', 'affec.tv', 'agkn.com', 'amazon-adsystem.com', 'amazonaws.com', 'analyticssystems.net', 'appier.net', 'aralego.com', 'aspnetcdn.com', 'audiencemanager.de', 'audrte.com', 'avct.cloud', 'azureedge.net', 'backblaze.com', 'basis.net', 'betrad.com', 'betweendigital.com', 'bidr.io', 'bidswitch.net', 'bidtheatre.com', 'bing.com', 'bkrtx.com', 'bliink.io', 'blismedia.com', 'bluekai.com', 'bluevoox.com', 'bnmla.com', 'boldchat.com', 'brand-display.com', 'bttrack.com', 'bumlam.com', 'casalemedia.com', 'chocolateplatform.com', 'ck-ie.com', 'clarium.io', 'clean.gg', 'clickagy.com', 'clickcertain.com', 'cloudflare.com', 'cognitivlabs.com', 'colossusssp.com', 'company-target.com', 'connectwise.com', 'contextweb.com', 'cookieless-data.com', 'cox.com', 'cox.net', 'cpmstar.com', 'cpx.to', 'createjs.com', 'creative-serving.com', 'creativecdn.com', 'criteo.com', 'crwdcntrl.net', 'datablocks.net', 'de17a.com', 'deepintent.com', 'deliverimp.com', 'demdex.net', 'deployads.com', 'digicert.com', 'digitaleast.mobi', 'disqus.com', 'dmxleo.com', 'dotomi.com', 'doubleclick.net', 'doubleverify.com', 'dyntrk.com', 'e-planning.net', 'e-volution.ai', 'emxdgt.com', 'eqads.com', 'erne.co', 'everesttech.net', 'evidon.com', 'exelator.com', 'extend.tv', 'eyeota.net', 'ezoic.com', 'facebook.com', 'fastly.net', 'fiftyt.com', 'fksnk.com', 'flashtalking.com', 'fout.jp', 'fwmrm.net', 'gamoshi.io', 'getgo.com', 'getpublica.com', 'gfx.ms', 'google-analytics.com', 'google.com', 'googleapis.com', 'googlesyndication.com', 'googletagservices.com', 'googleusercontent.com', 'gotomeeting.com', 'gstatic.com', 'gumgum.com', 'gvt1.com', 'gvt2.com', 'hostedrmm.com', 'hp.com', 'ib-ibi.com', 'id5-sync.com', 'impact-ad.jp', 'impactify.media', 'imrworldwide.com', 'includemodal.com', 'infolinks.com', 'inmobi.com', 'insightexpressai.com', 'intel.com', 'intentiq.com', 'ipredictive.com', 'jivox.com', 'jixie.io', 'js7k.com', 'kargo.com', 'krxd.net', 'ladsp.com', 'lencr.org', 'liadm.com', 'lijit.com', 'linkedin.com', 'live.com', 'lkqd.net', 'lmgssp.com', 'loopme.me', 'marphezis.com', 'mathtag.com', 'media.net', 'media6degrees.com', 'mediawallahscript.com', 'mfadsrvr.com', 'mgid.com', 'microsoft.com', 'microsoftonline.com', 'microsoftusercontent.com', 'mmi360.net', 'moatads.com', 'mookie1.com', 'mpeasylink.com', 'mrtnsvr.com', 'msauth.net', 'msedge.net', 'msftauth.net', 'msftncsi.com', 'mxptint.net', 'narrative.io', 'newrelic.com', 'nextmillmedia.com', 'ninthdecimal.com', 'nr-data.net', 'nrich.ai', 'octillion.tv', 'office.com', 'office.net', 'office365.com', 'ojrq.net', 'oktacdn.com', 'omnitagjs.com', 'onaudience.com', 'onetag-sys.com', 'openx.net', 'outbrain.com', 'outlook.com', 'owneriq.net', 'pippio.com', 'playground.xyz', 'postrelease.com', 'pswec.com', 'pubmatic.com', 'quantserve.com', 'quantumdex.io', 'realestate.com.au', 'researchnow.com', 'resetdigital.co', 'reson8.com', 'retargetly.com', 'rfihub.com', 'richaudience.com', 'rkdms.com', 'rlcdn.com', 'rqtrk.eu', 'rtactivate.com', 'rtbsrv.com', 'rubiconproject.com', 'scorecardresearch.com', 'screenconnect.com', 'sectigo.com', 'semasio.net', 'servenobid.com', 'sharepoint.com', 'sharethrough.com', 'simpli.fi', 'sitescout.com', 'smaato.net', 'smadex.com', 'smartadserver.com', 'smartclip.net', 'smrtb.com', 'socdm.com', 'sojern.com', 'sonobi.com', 'soundcast.fm', 'sportradarserving.com', 'spotim.market', 'spotxchange.com', 'springserve.com', 'stackadapt.com', 'stickyadstv.com', 'storygize.net', 'streamtheworld.com', 'sundaysky.com', 'svc.ms', 'syncedtool.com', 'taboola.com', 'tapad.com', 'tappx.com', 'teads.tv', 'technoratimedia.com', 'thebrighttag.com', 'thrtle.com', 'tiqcdn.com', 'topsrvimp.com', 'tremorhub.com', 'tribalfusion.com', 'truoptik.com', 'trustarc.com', 'truste.com', 'turn.com', 'tynt.com', 'udmserve.net', 'unrulymedia.com', 'uplynk.com', 'userreport.com', 'vidoomy.com', 'visx.net', 'w55c.net', 'walkme.com', 'wayfair.com', 'windowsupdate.com', 'y-medialink.com', 'yahoo.co.jp', 'yahoo.com', 'yandex.ru', 'yieldmo.com', 'yimg.com', 'zemanta.com', 'zeotap.com'

HP printer

hpeprint.com, hp.com, hp10.us

Example 2 – Anomaly detected for a device fingerprinted incorrectly

- The first device is fingerprinted as HP printer but looking at Wireshark its mac address is marked with “Universal Global Scientific Industrial Co., Ltd”
- Second device is an HP printer

Why is first device marked as anomaly ?

Median number of total queries

Anomaly device – 1900

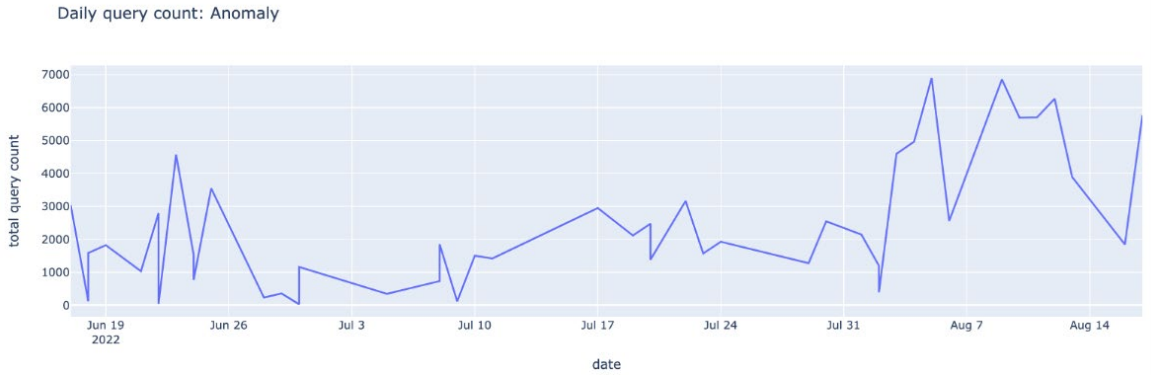
HP Printer - 55

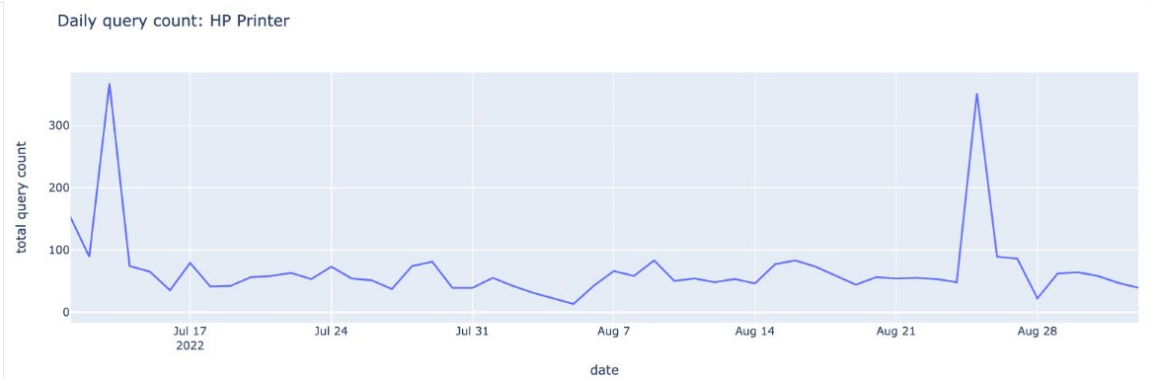
Median number of unique domains

Anomaly device – 64

HP Printer - 3

The domains queried by the anomalous device is too varied when comparing with most of the HP printers in the same customer





Conclusions

- DNS activity of a device can be used to identify device types.
- Passive DNS monitoring enable us to identify anomalous behavior of devices.
- Deep learning techniques can be used both for embedding DNS queries and can have several applications in the context of anomaly detection, application discovery and device classification.

References

[1] Efficient Estimation of Word Representations in Vector Space. Tomas Mikolov. Kai Chen, Gref Corrado, Heffrey Dean. 2013.
[2] Dns2Vec: Exploring Internet Domain Names Through Deep Learning. Amit Arora, scainet 2019.
[3] Detection of DNS Traffic Anomalies in Large Networks, Milan Cermak, Pavel Celeda , Jan Vykopal, 2014.

‘


Thank You
Asante
Shukria
Dhanyavadagalu
Vinaka
köszönöm
Kiitos
Maake
Manana
Dankon
Matondo
감사합니다
Kam Sah Hammida
شکرا
Dank Je
Dankscheen
спасибо
Mauruuru
Biyan
Diolch i Chi
Terima Kasih
Taiku
Tack
Blagodaram
Ngiyabonga
Dziekuje
Chokrane
Grazie
Juspaxar
Arigato
Mochchakkeram
Gracias
Khap
Paldies
Tingki
நன்றி
धन्यवाद
cảm ơn bạn
Gratias Tibi
Ua Tsaug Rau Koj
Bedankt
Dakujem
Grazas
Di Ou Mèsi
Obrigado
Děkuji
Nirringrazzjak
Kia Ora
Kop Khun
ありがとう
Suksama
Hvala
Rahmat
Nuwun
谢
Welalin
Djiere Dieuf
Matur
Misaotra
хвала
Danke
Merci
Go Raibh Maith Agat
Eskerrik Asko
Salamat
ขอขอบคุณคุณ
Najis Tuke