Cisco Umbrella

Identification of Malicious SSL Networks by Subgraph Anomaly Detection

Thomas Mathew, Dhia Mahjoub

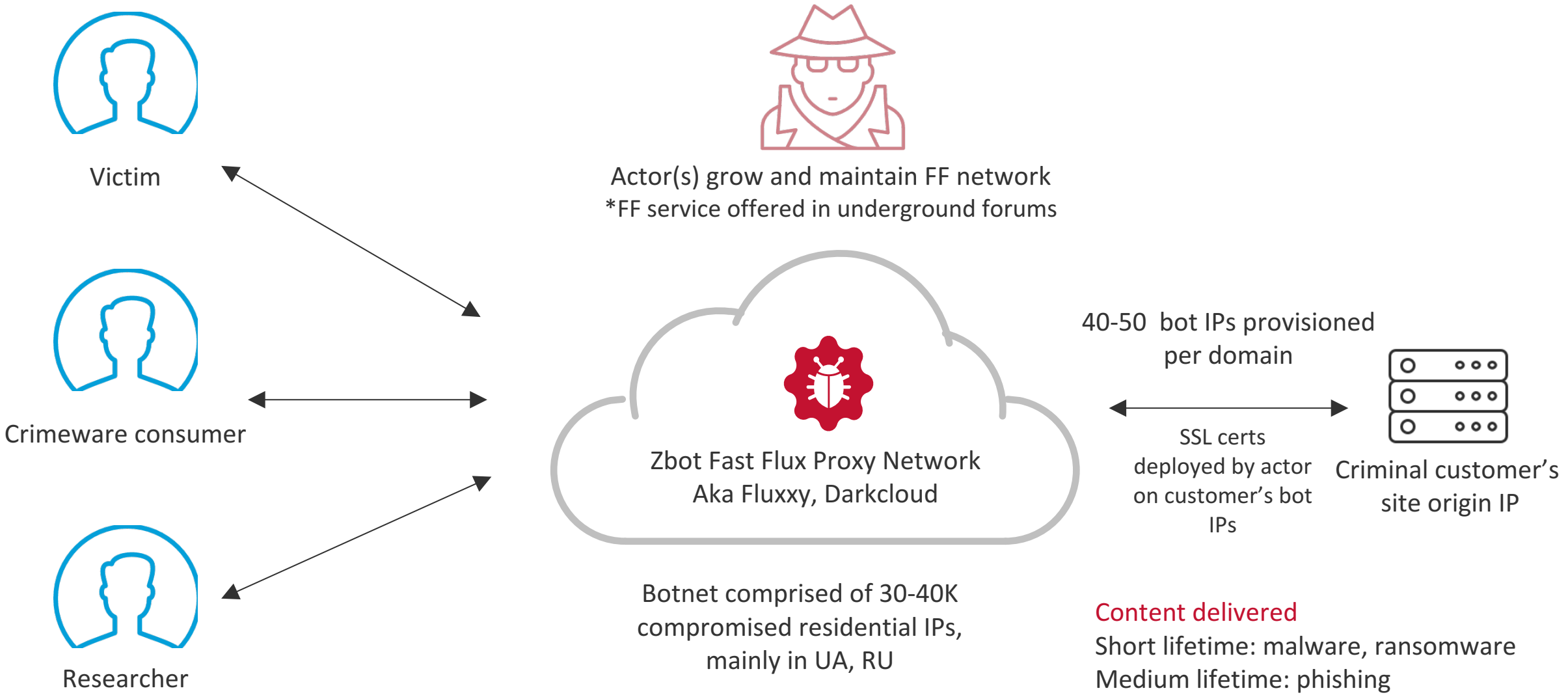
Tucson, Jan 10th 2018

Thomas Mathew, Senior Security Researcher, Cisco Umbrella (OpenDNS)
Data Science, Speaker at BruCon, Flocon, Kaspersky SAS, Black Hat, Defcon

Dhia Mahjoub, Head of Security Research, Cisco Umbrella (OpenDNS)
Graph theory, Speaker at Black Hat, Defcon, Flocon, RSA, Virus Bulletin
@DhiaLite

Zbot Fast Flux Bullet Proof Hosting operation

Introduced at Black Hat 2014, Botconf 2014, Defcon 2017

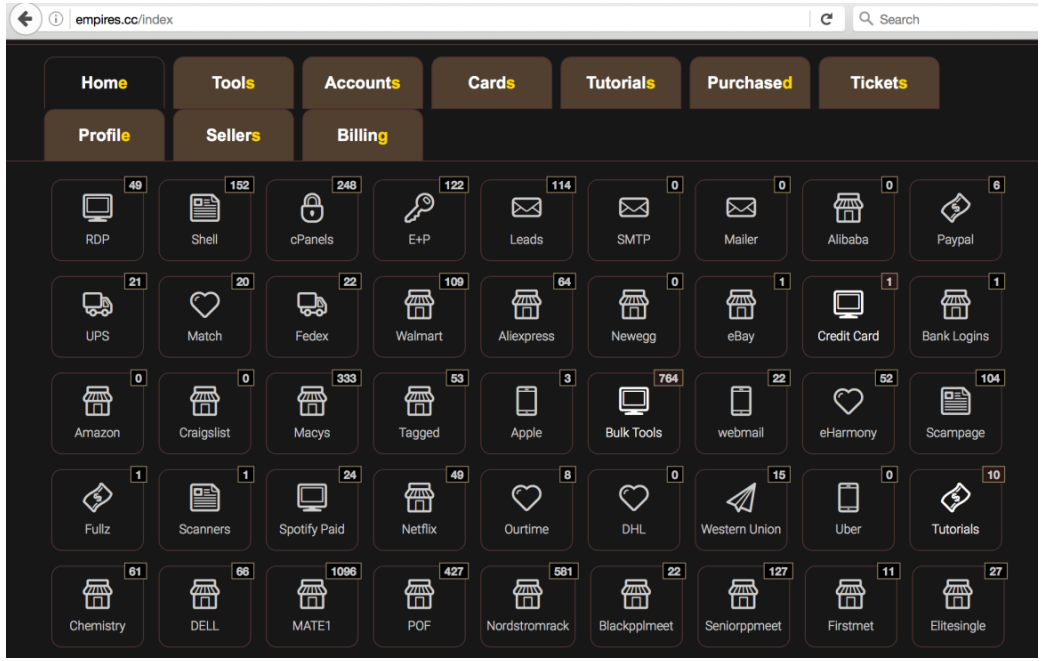


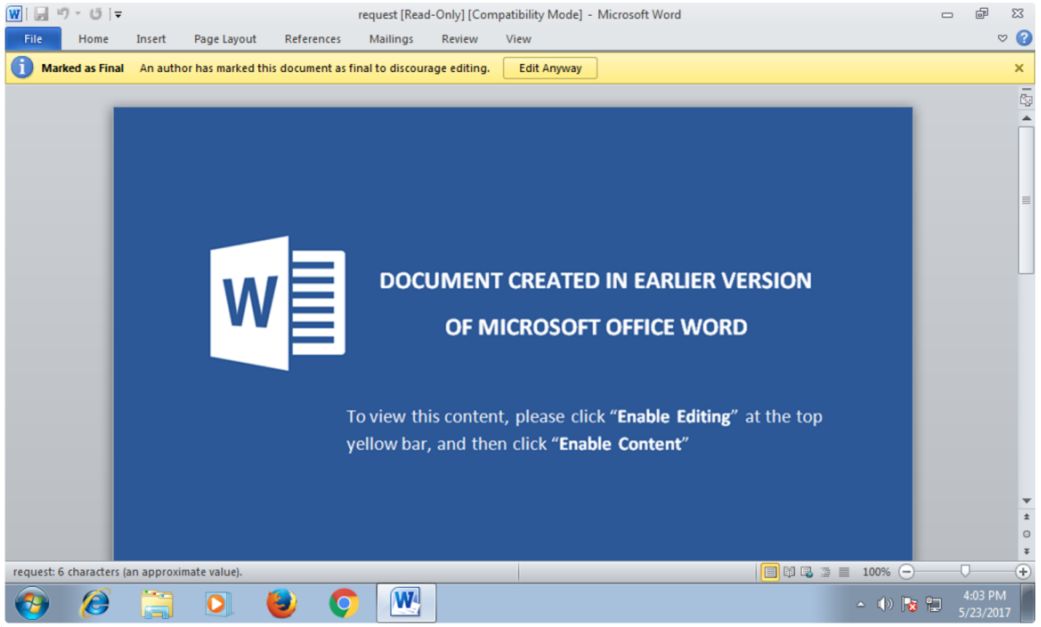
Content delivered

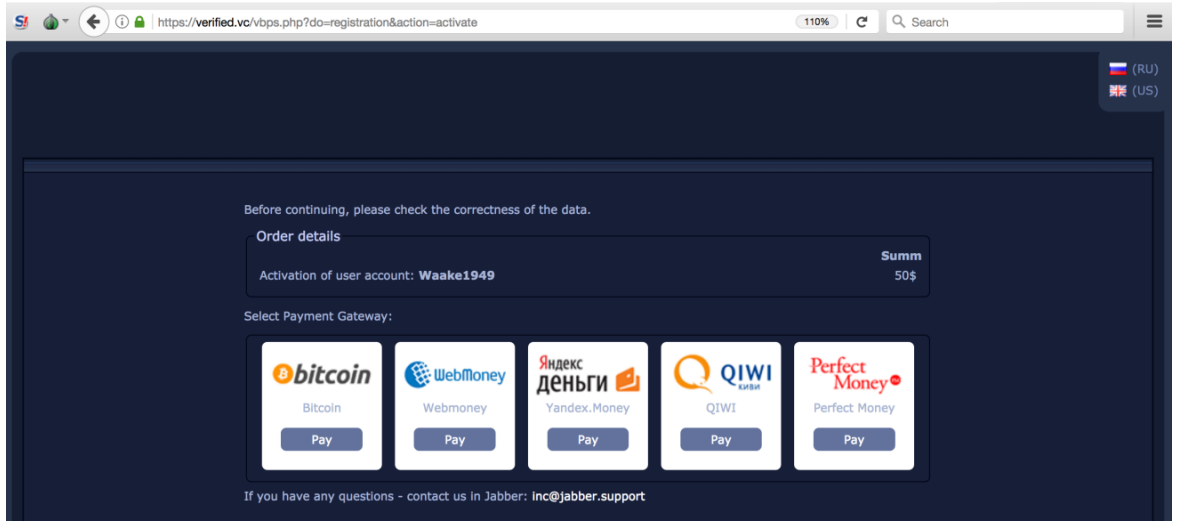
Short lifetime: malware, ransomware

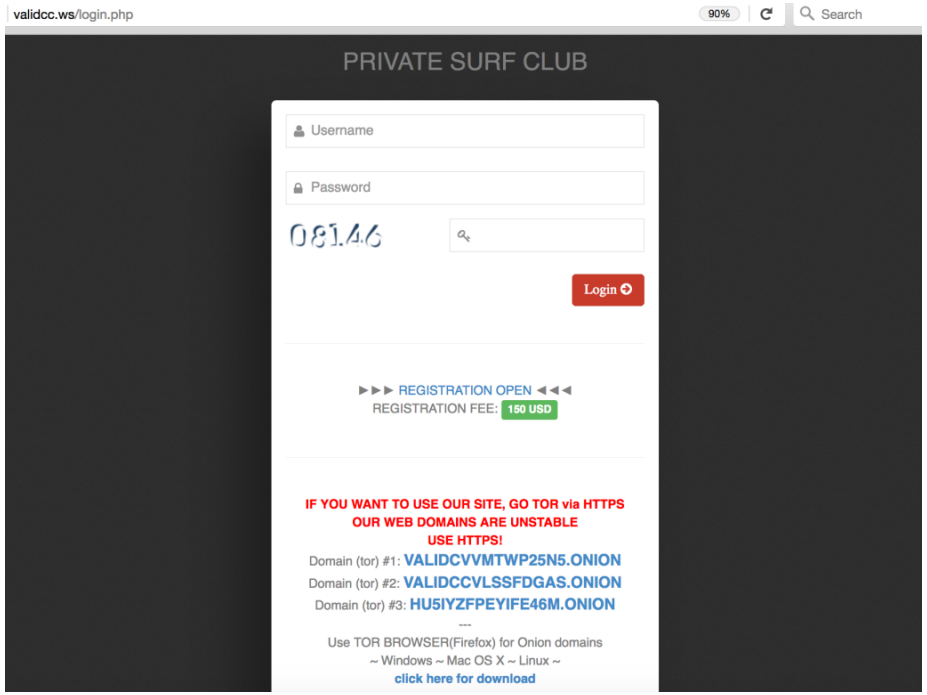
Medium lifetime: phishing

Long lifetime: carding, cybercrime forums


empires.cc/index
Search
Home
Tools
Accounts
Cards
Tutorials
Purchased
Tickets
Profile
Sellers
Billing
RDP 49
Shell 152
cPanels 248
E+P 122
Leads 114
SMTP 0
Mailer 0
Alibaba 0
Paypal 6
UPS 21
Match 20
Fedex 22
Walmart 109
Aliexpress 64
Newegg 0
eBay 1
Credit Card 1
Bank Logins 1
Amazon 0
Craigslist 0
Macys 333
Tagged 53
Apple 3
Bulk Tools 764
webmail 22
eHarmony 52
Scampage 104
Fullz 1
Scanners 1
Spotify Paid 24
Netflix 49
Ourtime 8
DHL 0
Western Union 15
Uber 0
Tutorials 10
Chemistry 61
DELL 66
MATE1 1096
POF 427
Nordstromrack 581
Blackpplmeet 22
Seniorppmeet 127
Firstmet 11
Elitesingle 27


request [Read-Only] [Compatibility Mode] - Microsoft Word
File
Home
Insert
Page Layout
References
Mailings
Review
View
Marked as Final
An author has marked this document as final to discourage editing.
Edit Anyway
DOCUMENT CREATED IN EARLIER VERSION
OF MICROSOFT OFFICE WORD
To view this content, please click "Enable Editing" at the top yellow bar, and then click "Enable Content"
request: 6 characters (an approximate value).
100%
4:03 PM
5/23/2017


https://verified.vc/vbps.php?do=registration&action=activate
110%
Search
(RU)
(US)
Before continuing, please check the correctness of the data.
Order details
Summ
Activation of user account: Waake1949
50$
Select Payment Gateway:
bitcoin
Bitcoin
Pay
WebMoney
Webmoney
Pay
Яндекс деньги
Yandex.Money
Pay
QIWI
QIWI
Pay
Perfect Money
Perfect Money
Pay
If you have any questions - contact us in Jabber: inc@jabber.support

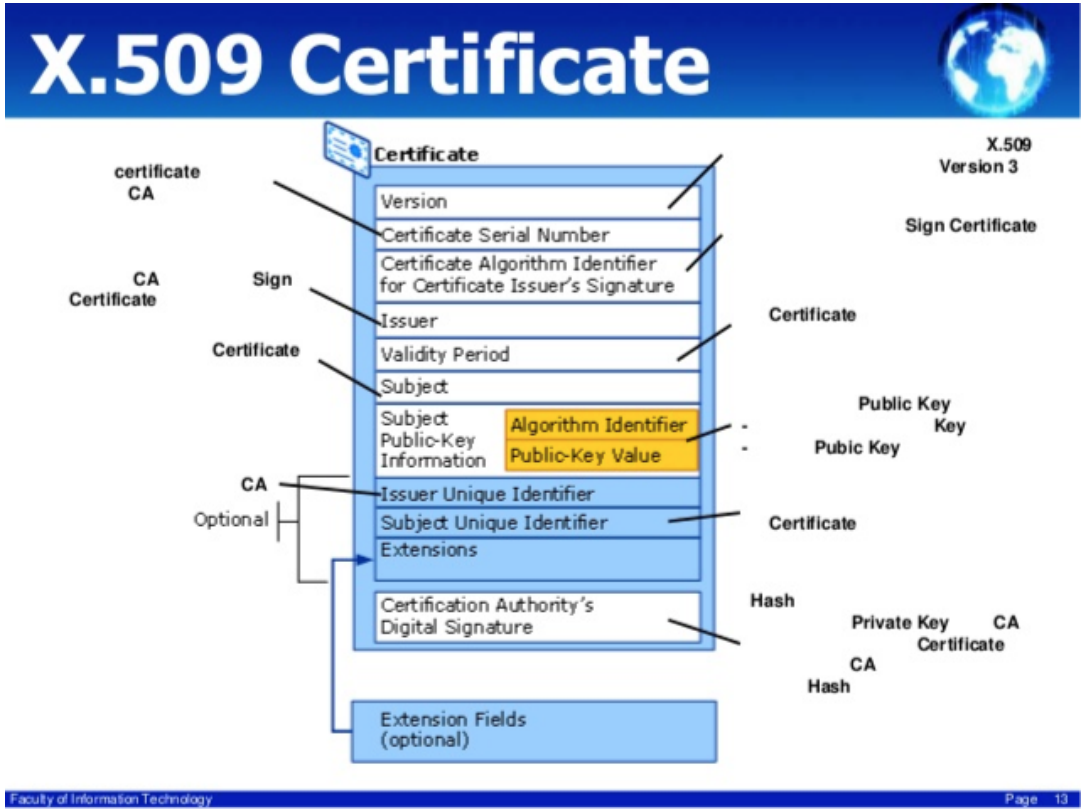

validcc.ws/login.php
90%
Search
PRIVATE SURF CLUB
Username
Password
081.46
Login
►►► REGISTRATION OPEN ◄◄◄
REGISTRATION FEE: 150 USD
IF YOU WANT TO USE OUR SITE, GO TOR via HTTPS
OUR WEB DOMAINS ARE UNSTABLE
USE HTTPS!
Domain (tor) #1: VALIDCVVMTWP25N5.ONION
Domain (tor) #2: VALIDCCVLSSFDGAS.ONION
Domain (tor) #3: HU5IYZFPEYIFE46M.ONION
Use TOR BROWSER(Firefox) for Onion domains
~ Windows ~ Mac OS X ~ Linux ~
click here for download

SSL CommonNames and Zbot FF

- Overall Goals:
  - Investigate the relationship between CommonNames found on x509 certificates and Zbot FF domains
  - Does Zbot FF's use of SSL 'leak information'?

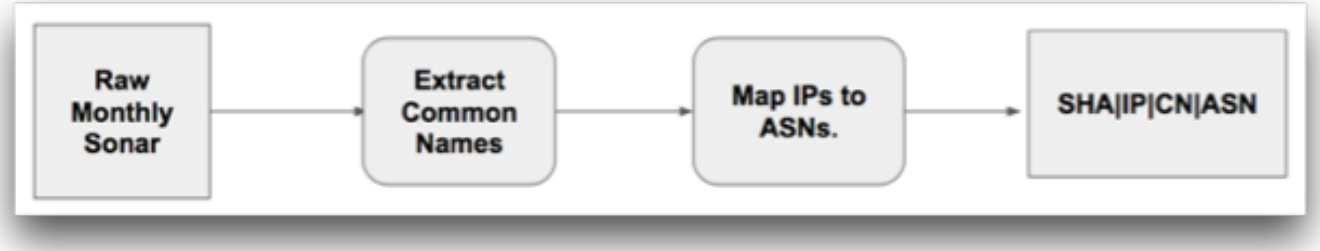
SSL - The Bare Minimum

- x509 certificates can give information about ownership of IP space
- Help track how a domain is hosted over a range of IPs/ASNs
- Identify relationships between common names and IP space



Sonar Data

- 2x or 4x monthly scans of IPv4 space minus networks who opt out
- Sonar SSL data contains information that allows us to map an x509 certificate to an IP that hosts the certificate
- Track Sonar Data over a 5 month period



Sonar Data

- Table documents the number of SHAs and CommonNames per month
- Manually inspecting the domains is infeasible

	Unique SHAs	Unique Common Names
JAN	1,068,402	850,236
FEB	692,542	589,609
MARCH	977,484	813,773
APRIL	249,252	233,834
MAY	1,098,914	958,321

Objectives

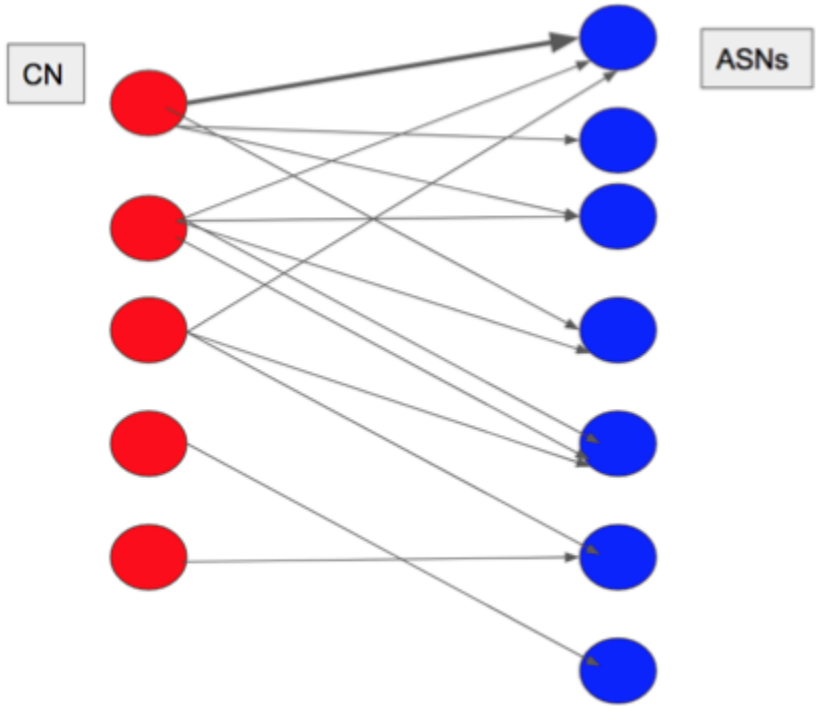
- Provide a series of simple statistical methods that allow a researcher to identify ZBot FF domains using SSL data
- More generally, provide a pipeline to analyze a large data set from a macro and micro view perspective to detect patterns

- All of the data discussed is open source and can be obtained at scans.io/study/sonar.ssl
- Thanks to rapid7 and the University of Michigan

Macro view analysis

Graphs as Representation

- SSL data naturally fits into a bipartite graph
- A bipartite graph is a graph whose vertices can be split into two disjoint sets
- Mapping could have been done differently:
  - IP range instead of ASN
  - ASN gave the best resolution to examine the data



Bipartite Graph - Investigation

- Anomalous behavior requires baseline metrics of normal
- Create a measure of popularity for each CommonName

Requirements:

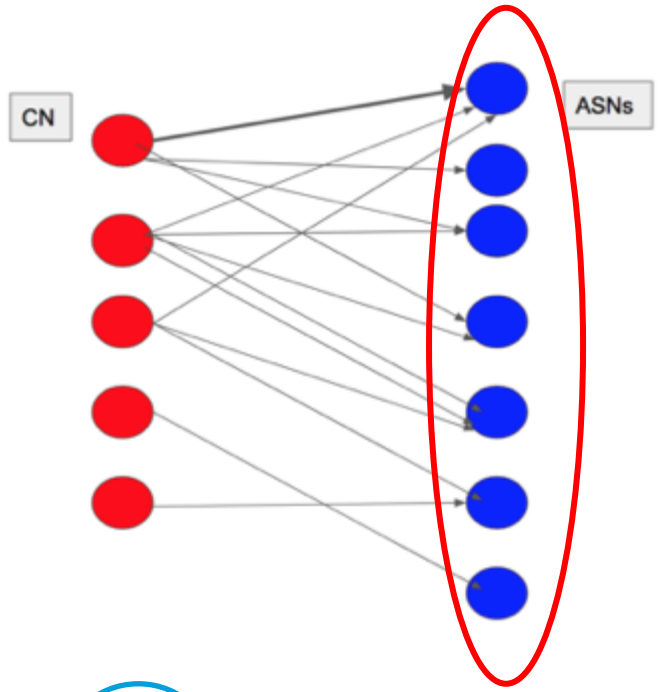
- Have it based on topological features of the bipartite graph
- Calculate frequencies of what type of ASNs host each CommonName
- Type of ASN refers to the popularity of the ASN
- Popularity of an ASN is defined by how many CommonNames are hosted on it

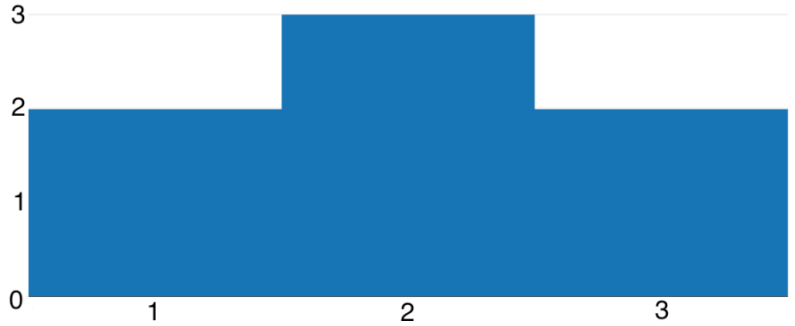
Graph Analytics - ASNs

- Examine the degree distributions for both sets of the bipartite graph

ASN degree

- ASNs with degree 1: 2
- ASNs with degree 2: 3
- ASNs with degree 3: 2

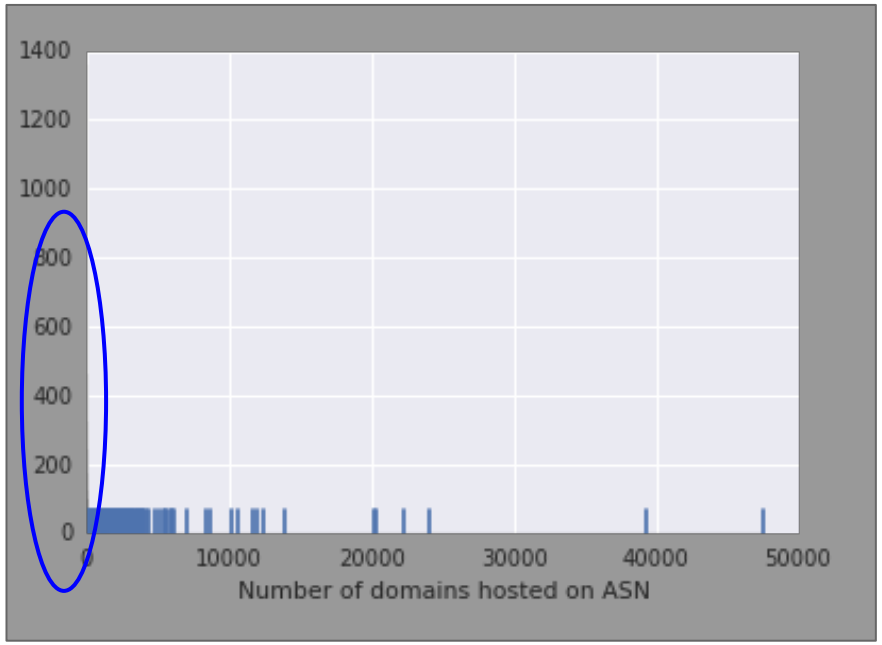




Graph Analytics - ASNs

- Sampled histogram shows popularity frequency of the 22,682 ASNs in the January SSL dataset from a new sample set of 5k

- Long tailed
  - There exists at least one ASN that hosts more than 50,000 unique CommonNames
  - Majority of mass concentrated in the range of ASNs hosting between 1 and 100 domains

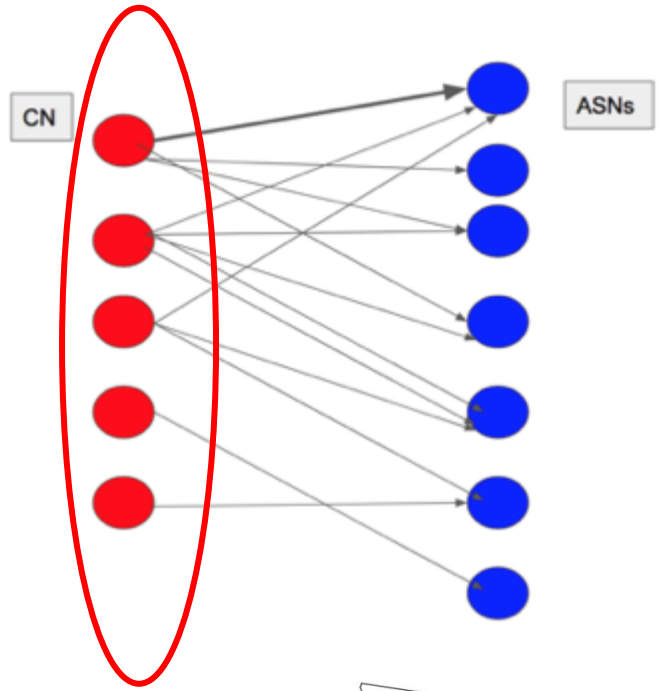


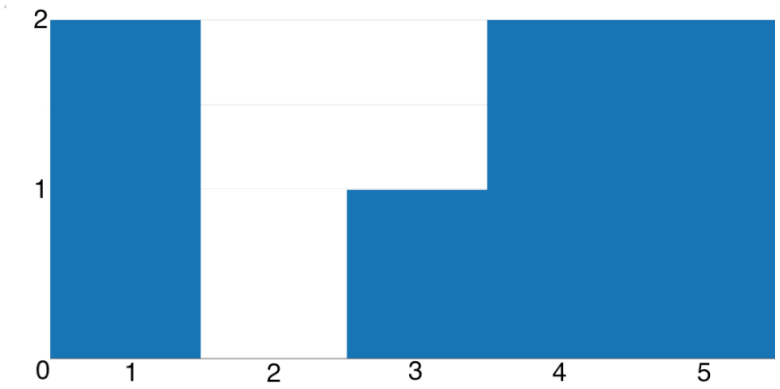
Graph Analytics - Domains

- Examine the degree distributions for both sets of the bipartite graph

CN degree

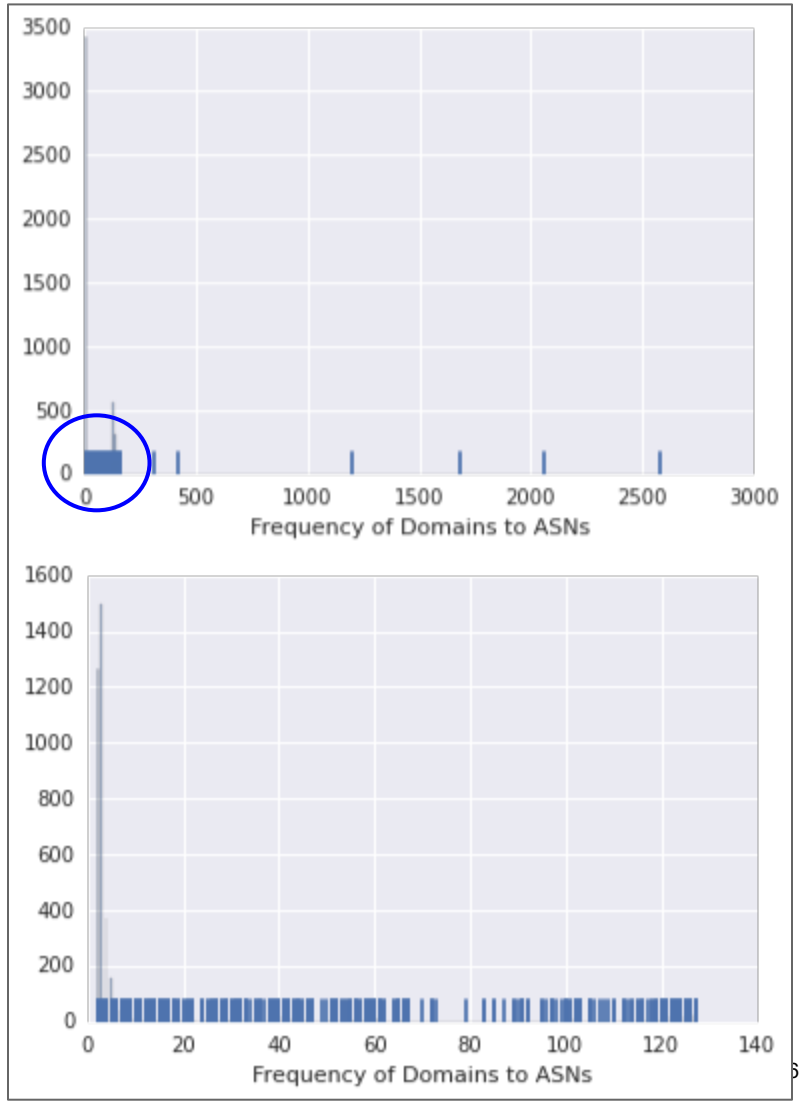
- CNs with degree 1: 2
- CNs with degree 2: 0
- CNs with degree 3: 1
- CNs with degree 4: 1
- CNs with degree 5: 1



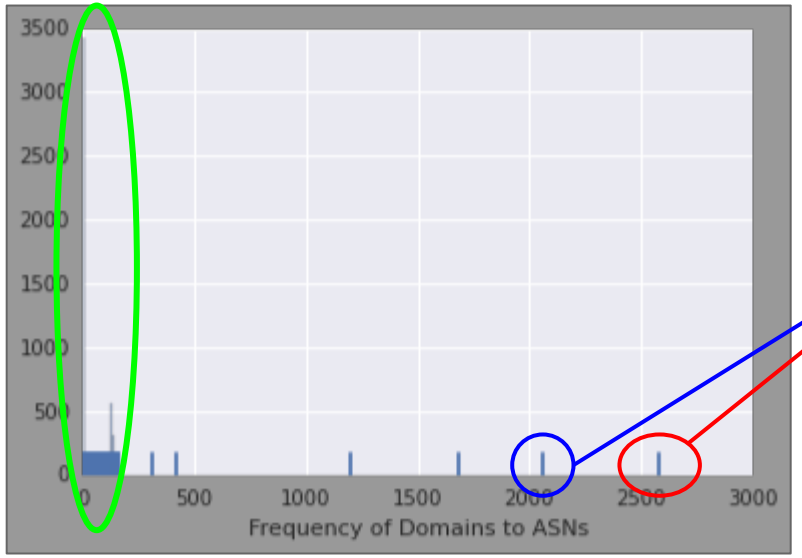


Graph Analytics - Domains

- First histogram displays a sampling from the set of 850,236 domains and their corresponding distribution
- The tick marks show the outliers at the tail
- Majority of the domains are hosted on (map to) 1-200 ASNs
- This is visually apparent by zooming into a sampled region of the the range 1-140



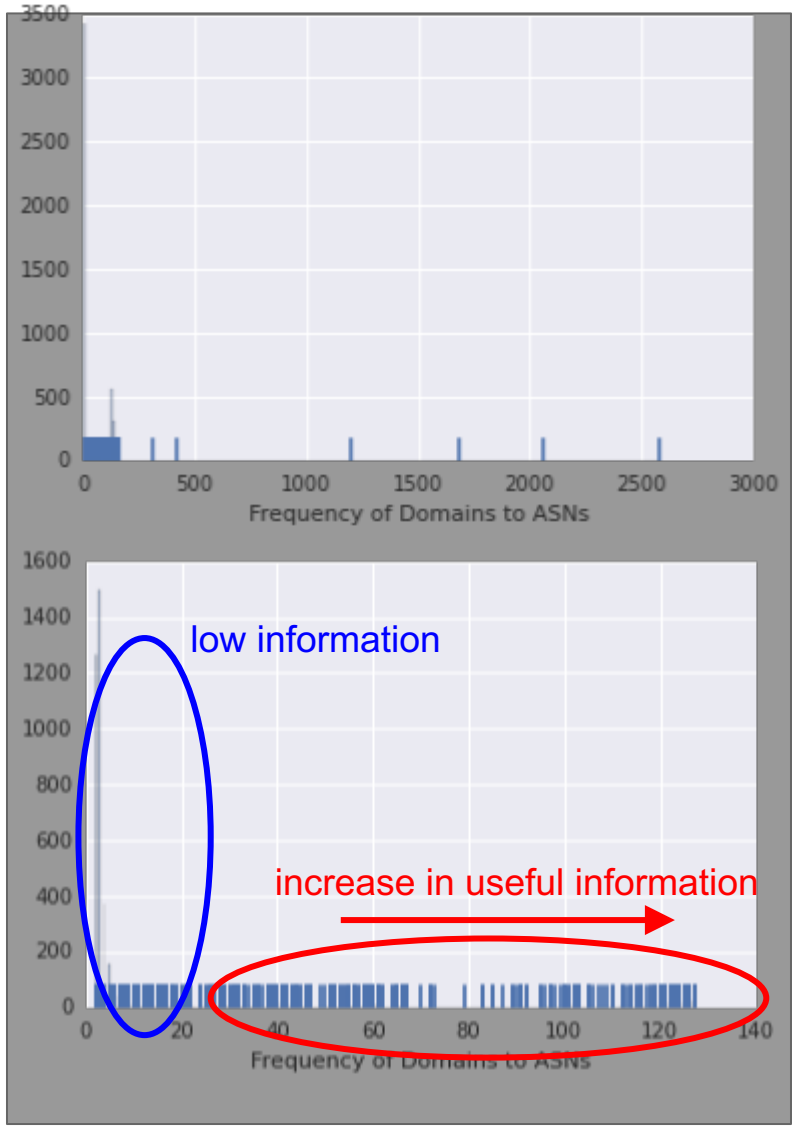
Graph Analytics - Domains



	domain	asn_count
0	www.dlink.com	2577
1	googlevideo.com	2060
2	google.com	1681
3	synology.com	1195
4	www.example.com	416
5	www.dlink.com.tw	311
6	itunes.apple.com	160
7	image-glb.qpyou.cn	159
8	asos-media.com	158
9	download.mcafee.com	157
10	www.koreanair.com	150

Graph Analytics - Domains

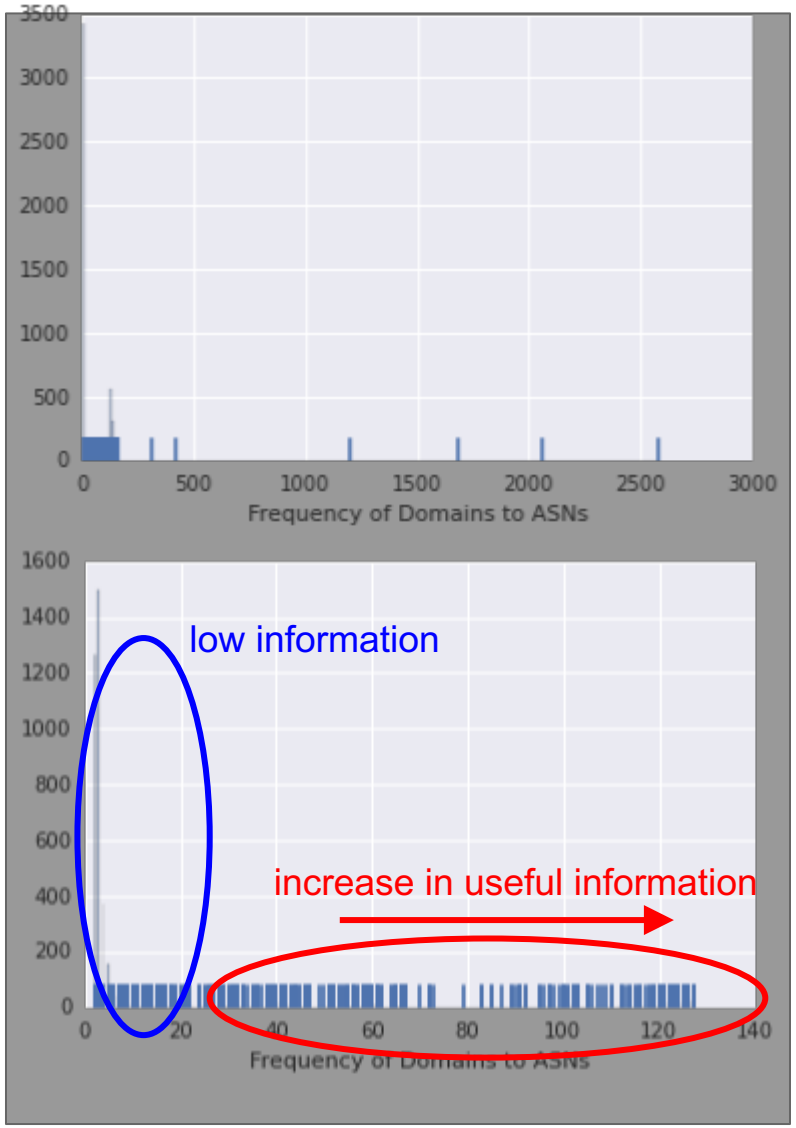
- As we move towards the right of the graph we ‘gain’ more information about the domain
  - Information regarding how popular it is
- Goal is to identify structures within the graph.
- Structure requires some measure of information



cisco

Graph Analytics - Domains

- 831,704 domains map to only a single ASN
  - translates to around 97% of all domains fall in that band
- 99% of domains map to the range between 1-10
- 1-10 too sparse → not enough information

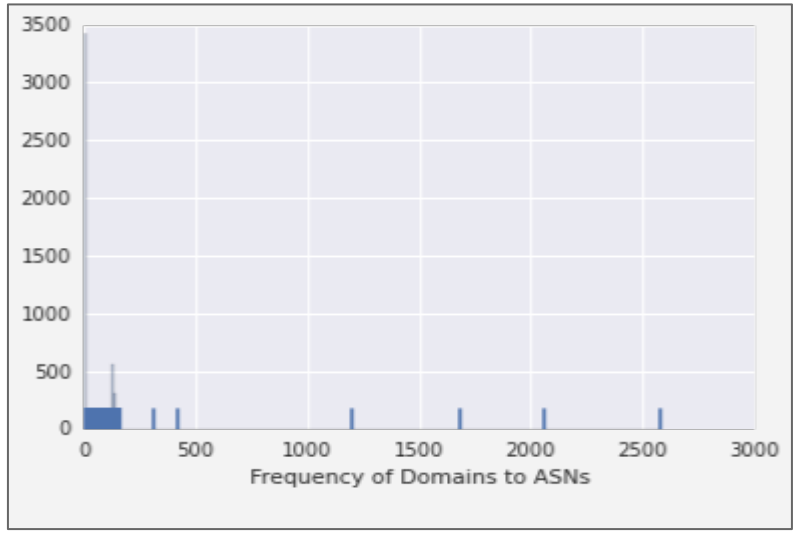


cisco

Graph Analytics - Domains

- Inspecting the histogram helps us determine an easy high pass filter
  - Use the mass density of the histogram in relation to information gain

- Examine domains that map to only 10 or more ASNs
  - Analogous to a document containing only one word
  - Allows us to remove around 99% of domains



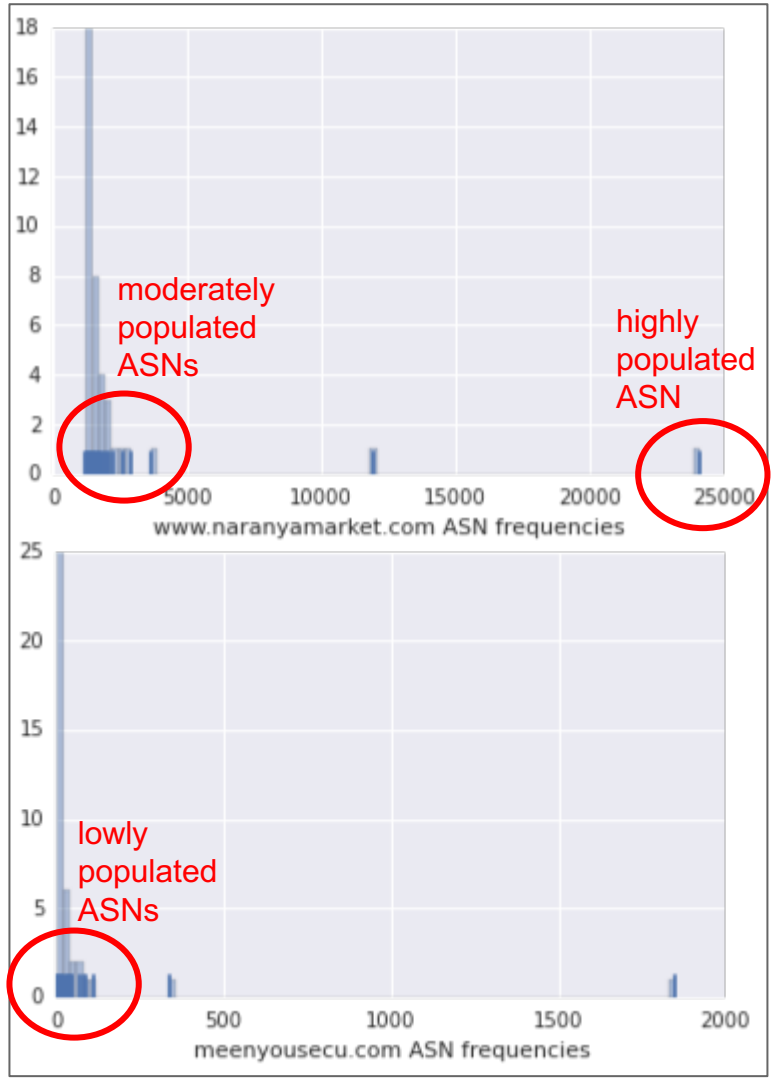
Micro view analysis

Individual Domain Analysis

- Taken the macro approach so far
  - Examined domain and ASN distribution

New Goal:

- Examine the ASN distribution of two individual domains
  - Histogram shows frequency of *different* types of ASN that hosts a cert for the given domain
  - X-axis denotes the type of ASN
  - Y-Axis frequency



Individual Domain Analysis

- Noticeable differences between the two domains
- naranyamarket.com is found on some extremely populated ASNs
  - One ASN that hosts ~ 25k domains
- meenyousecu.com max ASN hosts ~ 1.8k domains
- For meenyousecu.com - majority of ASNs host less than 108 unique CNs

```
In [42]: 'meenyousecu.com'

         [1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 9, 10, 11, 12, 15, 17, 17, 23, 28, 32, 36, 38, 40, 46, 48, 70, 7
         9, 88, 108, 339, 1848]
```

```
In [49]: 'www.naranyamarket.com'

         [1208, 1255, 1271, 1277, 1307, 1317, 1349, 1388, 1391, 1391, 1397, 1397, 1398, 1404, 1416, 1422, 1424, 1448, 1458, 1478
         , 1489, 1506, 1508, 1516, 1631, 1636, 1697, 1715, 1787, 1825, 1956, 1965, 2066, 2247, 2607, 2879, 3638, 11947, 24111]
```

Filter Design

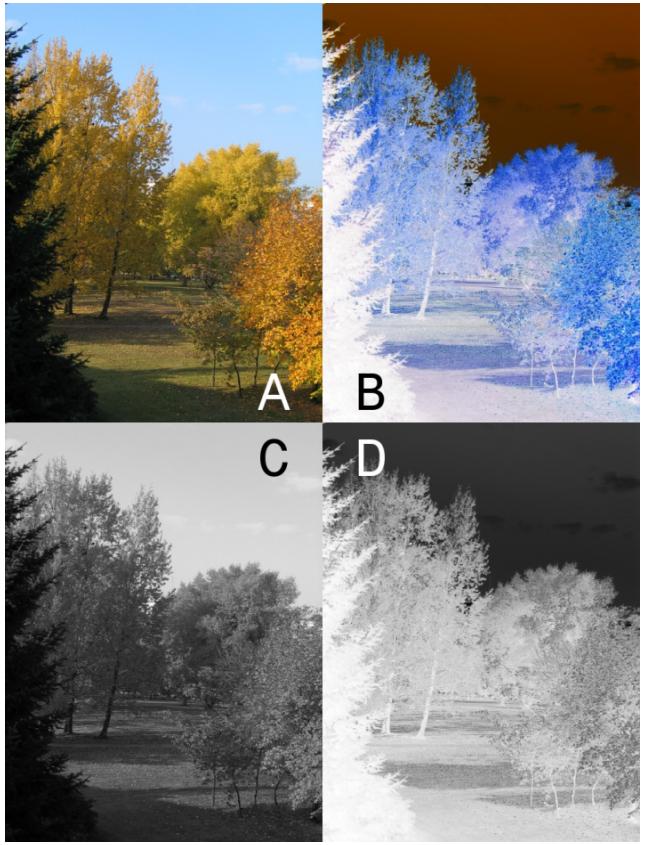
- You cannot directly compare histograms
  - Need to be compared on the same scale

- Devise a bucketing scheme that is sensitive to ‘low popularity’ ASNs
  - Increase the number of bands in the lower frequency spectrum
  - lower frequency spectrum refers to ASNs that map to <50 domains

- Intuition is based on the original distribution of ASN frequency counts
  - Interested in domains that are found on many ASNs where those ASNs are not highly popular
  - Majority of the domains hosted on 1-200 ASNs

Filter Design

- Bucket the frequencies into 9 different bands.
  - 1-5
  - 6-10
  - 11-20
  - 21-50
  - etc
  - Distance between the bands increase as we increase the ASN popularity
- Lower frequencies provide higher resolution
- The bucketing process transforms the distribution of ASN popularity counts into a 9-d vector



Filter Design - Histograms to Vectors

Different bands are: [1-5], [6-10], [11,20], [21,50], etc

Histogram

```
In [42]: 'meenyousecu.com'

[1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 9, 10, 11, 12, 15, 17, 17, 23, 28, 32, 36, 38, 40, 46, 48, 70, 7
9, 88, 108, 339, 1848]
```

Vector

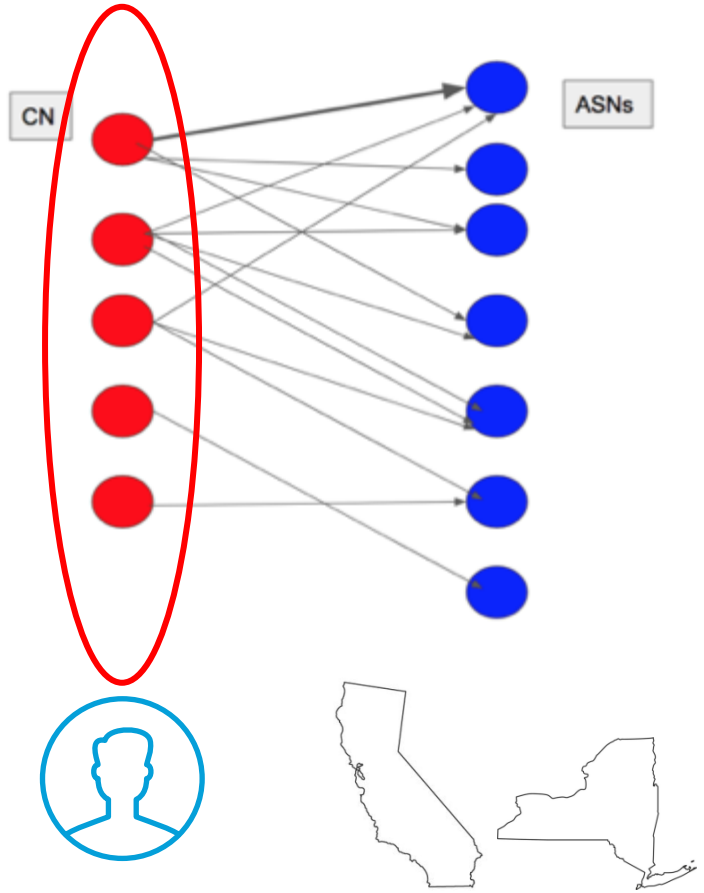
```
In [52]: v2 = create_hist_vect('meenyousecu.com')
         print v2

[ 15.  5.  5.  8.  3.  2.  0.  1.  0.]
```

```
In [49]: 'www.naranyamarket.com'

[1208, 1255, 1271, 1277, 1307, 1317, 1349, 1388, 1391, 1391, 1397, 1397, 1398, 1404, 1416, 1422, 1424, 1448, 1458, 1478
, 1489, 1506, 1508, 1516, 1631, 1636, 1697, 1715, 1787, 1825, 1956, 1965, 2066, 2247, 2607, 2879, 3638, 11947, 24111]
```

```
In [53]: v2 = create_hist_vect('www.naranyamarket.com')
         print v2

[ 0.  0.  0.  0.  0.  0.  0. 32.  7.]
```

Anomaly Detection per ASN Count Band

- Refocus on CNs and how many ASNs each domain maps to



Anomaly Detection per ASN Count Band

- Filter out the outliers (dlink, google, etc)
- Bucket domains that are mapped to roughly the same number of ASNs together
  - e.g. [160-150] → [itunes.apple.com, image-glb.qpyou.cn, asos-media.com, etc.]
- For a given count band, e.g. [160-150] calculate the pairwise Euclidean distance between two domains using the domain's histogram vector

	domain	asn_count
0	www.dlink.com	2577
1	googlevideo.com	2060
2	google.com	1681
3	synology.com	1195
4	www.example.com	416
5	www.dlink.com.tw	311
6	itunes.apple.com	160
7	image-glb.qpyou.cn	159
8	asos-media.com	158
9	download.mcafee.com	157
10	www.koreanair.com	150

[160, 150] band

Anomaly Detection per Frequency Band

- Hypothetical distance matrix for a band containing 3 domains
- Each column gives us the distance between a domain and every other domain in that band
  - i.e. red → all the distances from d1

	d1	d2	d3
d1	0	d(d1, d2)	d(d1,d3)
d2	d(d2, d1)	0	d(d2, d3)
d3	d(d3,d1)	d(d3,d2)	0

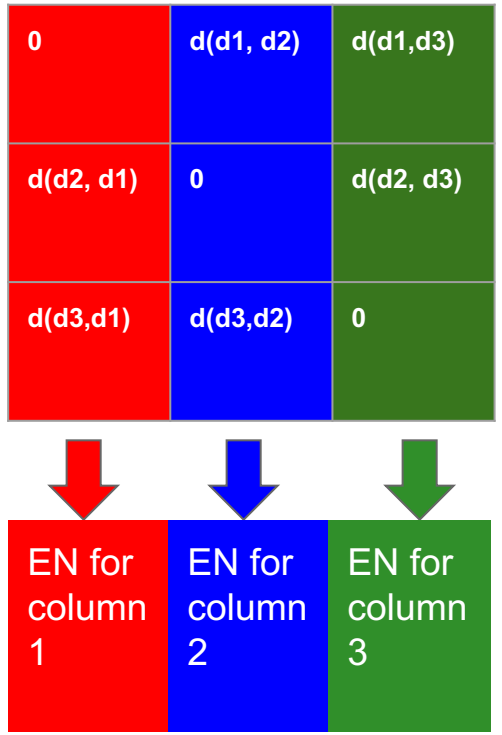
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Anomaly Detection per Frequency Band

- Calculate the Euclidean norm (EN) per column to determine the similarity of each domain to others

- Ran this algorithm over the Jan dataset.
  - Returned one interesting case in the band 100-110
  - e.g. EN(tangerine-secure.com) = 567.003



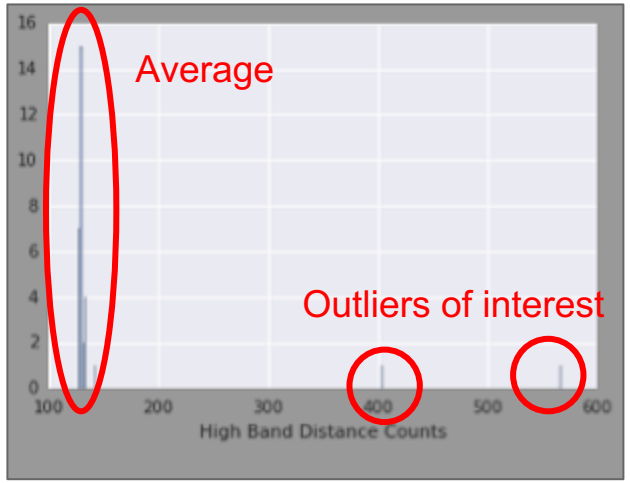

```
Out[94]: [('leonisavirtual.leonisa.com', 141.70744511139841),
          ('uis.uat.usajobs.gov', 132.38579984273238),
          ('frswebservices.org', 133.77966960640916),
          ('clarovideo.net', 133.31916591398252),
          ('tangerine-secure.com', 567.00352732588885),
          ('mobileonline.td.com', 131.44200241931802),
          ('www.getnet.com.br', 133.09019498069722),
          ('sanssl-014.bsdtools.com', 131.67004215082488),
          ('www.toysrus.de', 130.00384609695206),
          ('www.toysrus.fr', 130.00384609695206),
          ('cert2.coxmediagroup.com', 129.0310040261642),
          ('www.toysrus.at', 130.00384609695206),
          ('api.services.westjet.com', 130.00384609695206),
          ('stage.ritzcarlton.com', 130.00384609695206),
          ('www.brp.com', 128.7944098165755),
          ('edge-cdn.net', 129.73048986263791),
          ('midatlantic.aaa.com', 128.05467582247826),
          ('webssl.chinanetcenter.com', 403.43277011170028),
          ('tickets.cirquedusoleil.com', 128.7944098165755),
          ('www.santander.cl', 128.7944098165755),
          ('www.borbonese.com', 128.7944098165755),
          ('secure.boulanger.fr', 128.7944098165755),
          ('media1.1800flowers.com', 128.7944098165755),
          ('stage.brighttalk.net', 128.7944098165755),
          ('www.atgstores.com', 127.81627439414747),
          ('blueapron.com', 129.77287852244012),
```

Anomaly Detection per Frequency Band

- The histogram clearly shows a significant outlier (more than 2 std away)

  The outlier in the high band was ‘tangerine-secure.com’ which we verified as a ZBot Fast flux domain

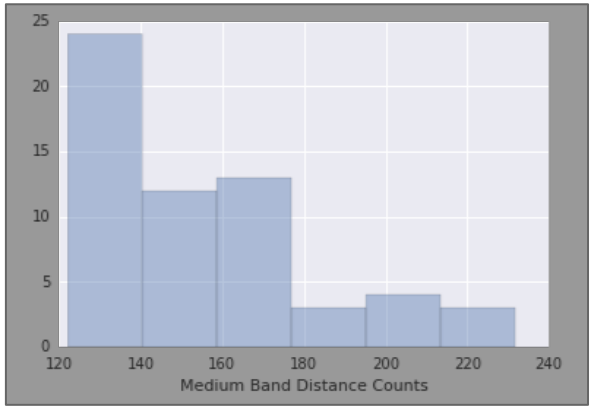
```
Out[94]: [('leonisavirtual.leonisa.com', 141.70744511139841),
          ('uis.uat.usajobs.gov', 132.38579984273238),
          ('frswebservices.org', 133.77966960640916),
          ('clarovideo.net', 133.31916591398252),
          ('tangerine-secure.com', 567.00352732588885),
          ('mobileonline.td.com', 131.44200241931802),
          ('www.getnet.com.br', 133.09019498069722),
          ('sanssl-014.bsdtools.com', 131.67004215082488),
          ('www.toysrus.de', 130.00384609695206),
          ('www.toysrus.fr', 130.00384609695206),
          ('cert2.coxmediagroup.com', 129.0310040261642),
          ('www.toysrus.at', 130.00384609695206),
          ('api.services.westjet.com', 130.00384609695206),
          ('stage.ritzcarlton.com', 130.00384609695206),
          ('www.brp.com', 128.7944098165755),
          ('edge-cdn.net', 129.73048986263791),
          ('midatlantic.aaa.com', 128.05467582247826),
          ('webssl.chinanetcenter.com', 403.43277011170028),
          ('tickets.cirquedusoleil.com', 128.7944098165755),
          ('www.santander.cl', 128.7944098165755),
          ('www.borbonese.com', 128.7944098165755),
          ('secure.boulanger.fr', 128.7944098165755),
          ('medial.1800flowers.com', 128.7944098165755),
          ('stage.brighttalk.net', 128.7944098165755),
          ('www.atgstores.com', 127.81627439414747),
          ('blueapron.com', 129.77287852244012),
```



Anomaly Detection per Frequency Band

- Another interesting band is the 30-40 ASN popularity range
- The distance spectrum is tighter than the 100-110 range
- The tail end of the histogram contains interesting domains

```
Out[104]: [('casino.netbet.com', 154.65445354078881),
           ('r-ak.bstatic.com', 160.09996876951601),
           ('meenyousecu.com', 231.53401477968632),
           ('www.naranyamarket.com', 142.09855734665288),
           ('swf.dougaku.tv', 160.09996876951601),
           ('www.wondershare.com', 154.65445354078881),
           ('realvu.net', 154.83216720048841),
           ('scdn1.ssl.altcdn.com', 154.83216720048841),
           ('secure.dot-st.com', 154.83216720048841),
           ('ustat.petpic.jp', 154.83216720048841),
           ('test-p-drv-dsl.support.ricoh.com', 154.83216720048841),
           ('www.exceda.com.ar', 141.04254677224174),
           ('www.dom.daimler.com', 141.04254677224174),
           ('js.rtoaster.jp', 154.83216720048841),
           ('cdn.zamimg.com', 154.83216720048841),
           ('bingowithkerry.com', 133.92535234226565),
           ('support3.cdnetworks.net', 166.99401186868948),
           ('support.cdnetworks.net', 173.88789492083686),
           ('fabulousbingo.co.uk', 128.7206277175496),
           ('ssl.cdngc.net', 165.14539048971363),
           ('sunbingo.co.uk', 127.09051892253804),
           ('support2.cdnetworks.net', 165.99397579430405),
           ('galacasino.com', 127.09051892253804),
           ('www.infoblox.com', 189.01058171435798),
           ('bdydns.com', 210.09521650908667),
```



Anomaly Detection - Results

Out of these 5 domains - three are ZBot fast flux (meenyousecu.com, securedatassl.net, secure.tangerineaccess.com

- Confirmed TPs via PDNS and active probing (of SSL certs)
- Removed FPs with additional signals
  This anomaly detection method was able to reduce ~800k domains down to a manageable list of 8 domains

```
meenyousecu.com:231.53401478
bdydns.com:210.095216509
securedatassl.net:206.300751332
secure.tangerineaccess.com:214.35717856
flxdns.com:204.514058196
```

Validating TPs and Removing FPs

- Further signals were incorporated to help delineate between outliers
- Examine the number of unique SHAs associated with a particular commonName
- Examine ASN count over IP count ratio
  - ZBot FF is bounded by the number of IPs the actor provisions to client domains
  - For ZBot FF, on average a client domain gets 1 IP per ASN
- Gave us actionable intelligence regarding which ASNs to monitor more closely

```
[('meenyousecu.com', 231.53401477968632),
 ('example.com', 219.68613975396809),
 ('secure.tangerineaccess.com', 214.35717855952481),
 ('bdydns.com', 210.09521650908667),
 ('securedatassl.net', 206.30075133164203),
 ('flxdns.com', 204.51405819649659),
 ('51cto.com', 200.31724838365767),
 ('www.infoblox.com', 189.01058171435798),
```

```
In [70]: test_data2[test_data2.apply(lambda x: x['dist'] > 185, axis=1)]
```

Out[70]:

	domain	asn_count	ip_count	sha_count	ratio	dist
1393	example.com	39	450	288	11.538462	219.686140
1394	meenyousecu.com	39	51	1	1.307692	231.534015
1416	www.infoblox.com	34	130	56	3.823529	189.010582
1417	bdydns.com	33	461	1	13.969697	210.095217
1421	securedatassl.net	33	44	1	1.333333	206.300751
1422	www.ansible.com	32	169	95	5.281250	187.122954
1429	secure.tangerineaccess.com	32	37	1	1.156250	214.357179
1444	flxdns.com	31	326	1	10.516129	204.514058
1448	51cto.com	30	737	1	24.566667	200.317248

ASNs of ZBot FF vs Popular Domains

- CNs on the ASN popularity band [30,40]
- secure.tangerineaccess.com
  - AS50161|PE Vasylchyshyn Iurii Oleksandrovych
  - AS15895|Kyivstar PJSC
  - AS42975|Chiliy Valery Mykhaylovych PE

- blueapron.com
  - AS20940|Akamai International B.V.
  - AS16625|Akamai Technologies, Inc.
  - AS5511|Orange
  - AS14618|Amazon.com, Inc.

Takeaways

- Global structure of SSL graph can inform local behavior of SSL graph
  - We can use the macro knowledge to help us create similarity measures for each micro layer
- We can use this statistical method as a mechanism to cluster anomalous domains in the graph
  - We can reduce the set of domains of interest from ~800k → 8
- Unsupervised method to detect suspicious domains on IP/SSL space
- This method revealed malicious domains in other monthly scans:
  - mobilebanking1.scotiabankcan.com (Feb)
  - datasslsecure.net (Feb)

Our other related work

- Defcon 2017 https://www.youtube.com/watch?v=AbJCOVLQbjs
- Black Hat 2017
- Usenix Enigma 2017 https://www.youtube.com/watch?v=ep2gHQgjYTs&t=818s
- Black Hat 2016 https://www.youtube.com/watch?v=m9yqnwuqdSk
- RSA 2016 https://www.rsaconference.com/events/us16/agenda/sessions/2336/using-large-scale-data-to-provide-attacker
- BruCon 2015 https://www.youtube.com/watch?v=8edBgoHXnwg
- Virus Bulletin 2014 https://www.virusbtn.com/conference/vb2014/abstracts/Mahjoub.xml
- Black Hat 2014 https://www.youtube.com/watch?v=UG4ZUaWDXSs

Thank you

Questions?

We are hiring

Thomas Mathew, thomathe@cisco.com

Dhia Mahjoub, dmahjoub@cisco.com