Security in Communications Networks - P2

Introduction and Methodology

To catch anomalous activity in a network, it is first necessary to define what is non-anomalous behaviour. For this end, from a provided dataset (data5) containing the typical behaviour of a network's devices were extracted various metrics.

The dataset itself contains the following columns:

- Timestamp
- Source IP Address
- Destination IP Address
- Protocol Used
- Port Used
- Bytes Uploaded
- Bytes Downloaded

These were combined and analyzed in order to extract the following metrics:

- Number of packets per second, for each source/destination IP combination.
- Destination IP addresses and frequency, for each source IP..
- Geolocation of destination IPs, for each source IP, and frequency.
- Ports used for communications.
  - Ports used.
  - Volume of data, by port.
- Volume of data uploaded and downloaded, by source IP.
- Protocol used for communication.
- Visited domains that have uncommon TLDs (.xyz, .zip, .evil...). (Or a whitelist of common TLDs)

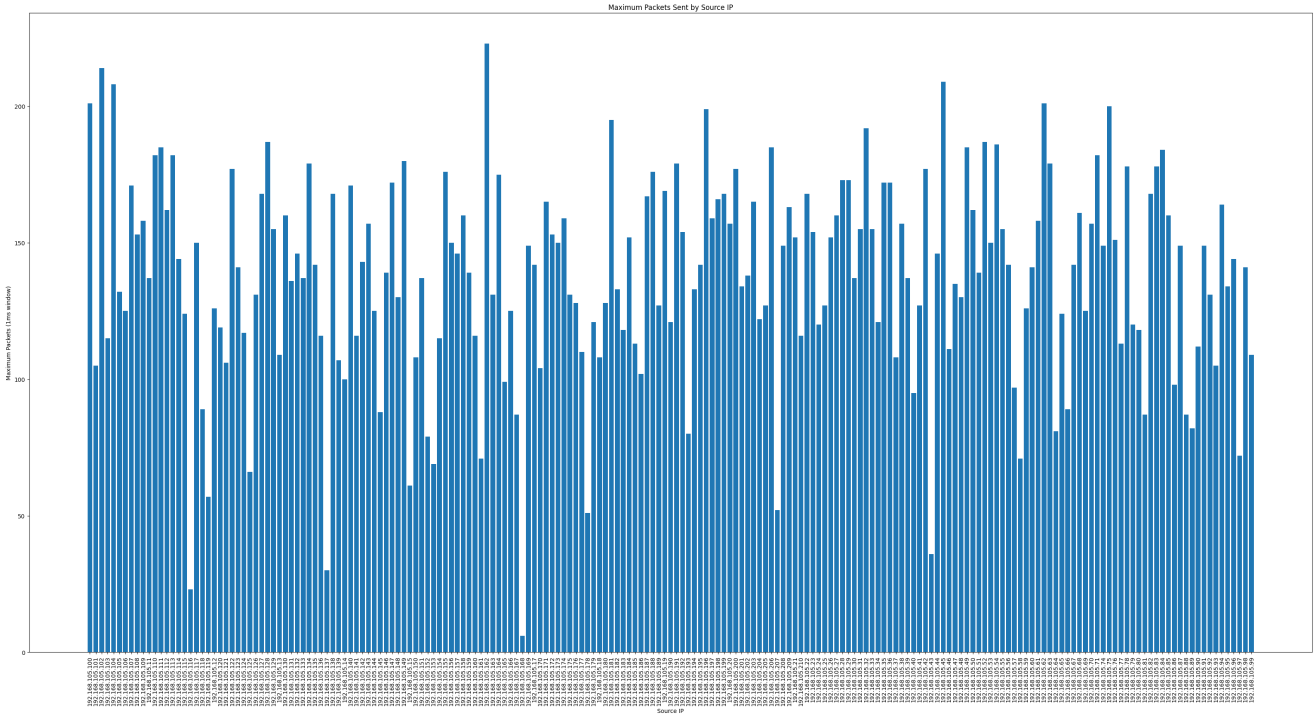
Analysis

Number of packets per second, for each source/destination IP combination.

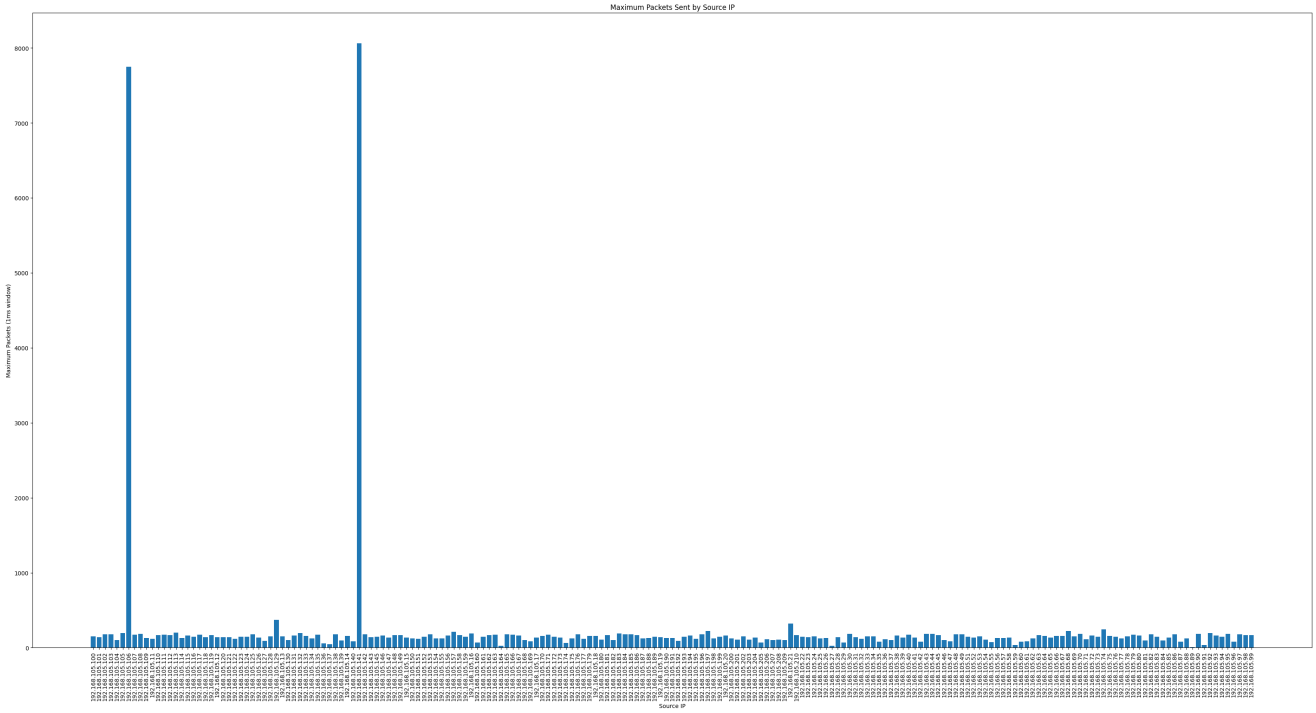
An exceedingly high frequency of packets, to either a single or a high amount of destination IPs can be a sign of an infected host carrying out a DDoS attempt.

Results

Non-anomalous analysis:



Anomalous analysis:



As it can be seen, there are two machines with very large spikes in the maximum frequency (packets in a single millisecond):

- 192.168.105.141
- 192.168.105.106

Spikes this high are most likely infected machines.
With the following data, extracted from the test data, we can see exactly where the source and destination of the packets. Important to note that both are internal IPs, which will be revisited later on.
(For comparison, the highest number outside of these high values is 372, and in the non-anomalous data, the highest is 175.)

```
       src_ip           dst_ip              packets
20339 192.168.105.141  192.168.105.225     8063
20342 192.168.105.141  192.168.105.238     7904
5222  192.168.105.106  192.168.105.225     7750
5225  192.168.105.106  192.168.105.238     7606
8021  192.168.105.141  192.168.105.225     6268
8024  192.168.105.141  192.168.105.238     6250
46178 192.168.105.141  192.168.105.238     5625
34859 192.168.105.141  192.168.105.238     5090
34858 192.168.105.141  192.168.105.225     4684
46175 192.168.105.141  192.168.105.225     4651
17360 192.168.105.106  192.168.105.225     2027
17363 192.168.105.106  192.168.105.238     1842
```

Detection: A value threshold that takes into account normal behaviour, such as 5X, where X is average/max frequency for a machine. This is given as an example, as the actual metric would have to be fine-tuned to avoid false positives or negatives.

On the test data, a threshold of 2*X, where X is the maximum packet frequency for a source IP was used.
Applied to the test data, , these are the flagged IPs, with respective values:

```
             src_ip  data_packets  test_packets
6   192.168.105.106          1000          7750
45  192.168.105.141           928          8063
72  192.168.105.168            48           102
```

To be noted that what's being measured here is the maximum amount of packets in a single milisecond, so, speed at which packets have been sent, and not overall number of packets.

Destination IP addresses and frequency, for each source IP

Communication with completely new IPs may be suspicious but not enough to be an immediate concern alone. It is just as likely to be an user accessing a new website as a malware contacting it's command-and-control.
Communications with local IP addresses may be suspicious, as an attempt to infect other devices in the network, if that machine doesn't often do so (Like an FTP server would, for instance).

Results

Note: No graphs were made for this metric, as the 3D graph needed to represent the data crashed the computer this work was done on.

Non-Anomalous Analysis:
(First rows)

	src_ip	dst_ip	packets	packets_percent
19760	192.168.105.206	192.168.105.225	872	0.089997
19680	192.168.105.206	142.250.200.68	850	0.087726
92	192.168.105.100	142.250.200.68	835	0.086178
19763	192.168.105.206	192.168.105.238	800	0.082566
14339	192.168.105.181	172.217.17.14	753	0.077715
29883	192.168.105.75	192.168.105.238	734	0.075754
19731	192.168.105.206	172.217.17.14	729	0.075238
29880	192.168.105.75	192.168.105.225	726	0.074928
18963	192.168.105.203	172.217.17.14	724	0.074722

Anomalous Analysis:
(First rows)

	src_ip	dst_ip	packets	packets_percent
8856	192.168.105.141	192.168.105.238	24869	2.419587
8853	192.168.105.141	192.168.105.225	23666	2.302543
1269	192.168.105.106	192.168.105.225	9777	0.951237
1272	192.168.105.106	192.168.105.238	9448	0.919227
6463	192.168.105.129	192.168.105.238	1190	0.115779
6460	192.168.105.129	192.168.105.225	1034	0.100601
7743	192.168.105.135	142.250.200.68	832	0.080948
33489	192.168.105.95	192.168.105.238	828	0.080559

```
20946   192.168.105.21  192.168.105.225      812         0.079002
```

Comparison:
Here, 'variation' is the change in % between test and non-anomalous data.
As there are 9071 rows (Combinations between source and destination IP addresses), changes of 0.1% and up are quite significant. In the non-anomalous data, the highest % a single row had was 0.08%, for a frame of reference.
Thus, the 2% increase present in the first two rows is a 25x increase relative to it. (Certainly anomalous.)

```
              src_ip           dst_ip  variation
2159  192.168.105.141  192.168.105.238   2.393269
2156  192.168.105.141  192.168.105.225   2.286133
326   192.168.105.106  192.168.105.225   0.928944
329   192.168.105.106  192.168.105.238   0.903746
1525  192.168.105.129  192.168.105.238   0.083269
5523   192.168.105.21  192.168.105.225   0.061147
1522  192.168.105.129  192.168.105.225   0.056325
8872   192.168.105.95  192.168.105.238   0.053003
1845  192.168.105.135   142.250.200.68   0.050502
1313  192.168.105.125  192.168.105.225   0.043197
7613   192.168.105.66  192.168.105.238   0.042298
```

Detection: Large variations in the % values of packets sent/received. Communication with a new IP, paired with other anomalies. Communications between two internal IPs that have never communicated before.

With a threshold of 0.1%, the following results would've been flagged:

```
              src_ip           dst_ip  variation
326   192.168.105.106  192.168.105.225   0.928944
329   192.168.105.106  192.168.105.238   0.903746
2156  192.168.105.141  192.168.105.225   2.286133
2159  192.168.105.141  192.168.105.238   2.393269
```

Geolocalization of destination IPs, for each source IP, and frequency.

Communication with IPs from a country previously not seen may be suspicious but not enough to be an immediate concern, depending on volume of data transferred.

Very large services such as google and AWS have their servers geographically spread out, and load-balancing on their side may mean a request can travel to a multitude of different countries. However, malware C2 servers, hosts for payloads, and other malicious websites are often hosted in countries with more lax laws regarding hosting, and those are often not used by "usual" services.

Results

Non-Anomalous data:
Percentage of destination addresses' countries:

```
US    35.657765
PT    29.651315
NL     1.892613
NA     1.363263
GB     0.843306
BR     0.579508
ES     0.215393
IE     0.127048
IN     0.056867
DE     0.048611
JP     0.038187
SG     0.035194
IT     0.026937
CA     0.025183
AU     0.023531
FR     0.023222
SE     0.022809
KR     0.018371
HK     0.018061
CN     0.014759
ZA     0.009908
IL     0.008979
CH     0.008979
BH     0.008257
AE     0.006089
SA     0.004748
```

```
NO     0.004644
ID     0.004231
MY     0.002683
TW     0.002374
BE     0.002374
AP     0.002271
CL     0.001445
PL     0.000929
DK     0.000103
```

Test data:

```
US    33.335117
PT    27.458699
NL     1.820358
NA     1.294585
GB     0.780779
BR     0.550291
ES     0.189625
IE     0.088731
RU     0.076473
CN     0.051565
JP     0.038139
DE     0.037166
AU     0.035901
IN     0.032301
SG     0.032107
HK     0.029091
IT     0.020918
FR     0.015567
SE     0.014983
CA     0.014594
KR     0.013621
ZA     0.013426
CH     0.012356
BH     0.005448
NO     0.004767
MY     0.004573
IL     0.004475
AE     0.004475
ID     0.002724
CL     0.002238
SA     0.002043
```

```
KG     0.001654
UA     0.001557
MM     0.001557
PL     0.001362
LB     0.001265
       0.001168
BE     0.000876
BY     0.000486
TW     0.000486
FI     0.000389
LU     0.000389
KH     0.000389
OM     0.000292
GE     0.000292
BA     0.000292
AT     0.000292
UZ     0.000292
SC     0.000292
TH     0.000195
EE     0.000097
NZ     0.000097
KZ     0.000097
AR     0.000097
```

Detection: New entries in the country list, paired with unusual volume of data.

New countries and variation:

```
idx country  variation  is_new
8       RU   0.076473   True
31      KG   0.001654   True
32      UA   0.001557   True
33      MM   0.001557   True
35      LB   0.001265   True
38      BY   0.000486   True
40      FI   0.000389   True
41      LU   0.000389   True
42      KH   0.000389   True
43      OM   0.000292   True
44      GE   0.000292   True
45      BA   0.000292   True
46      AT   0.000292   True
47      UZ   0.000292   True
48      SC   0.000292   True
```

```
49      TH   0.000195   True
50      EE   0.000097   True
51      NZ   0.000097   True
52      KZ   0.000097   True
53      AR   0.000097   True
9       CN   0.036807  False
12      AU   0.012370  False
15      HK   0.011029  False
21      ZA   0.003519  False
22      CH   0.003377  False
25      MY   0.001889  False
29      CL   0.000793  False
34      PL   0.000433  False
24      NO   0.000123  False
10      JP  -0.000048  False
37      BE  -0.001498  False
28      ID  -0.001507  False
27      AE  -0.001614  False
39      TW  -0.001887  False
30      SA  -0.002704  False
23      BH  -0.002808  False
14      SG  -0.003087  False
26      IL  -0.004504  False
20      KR  -0.004750  False
16      IT  -0.006019  False
17      FR  -0.007655  False
18      SE  -0.007826  False
19      CA  -0.010589  False
11      DE  -0.011445  False
13      IN  -0.024566  False
6       ES  -0.025769  False
5       BR  -0.029217  False
7       IE  -0.038317  False
4       GB  -0.062527  False
3       NA  -0.068679  False
2       NL  -0.072255  False
1       PT  -2.192616  False
0       US  -2.322648  False
```

With a threshold of 0.1% applied, the following countries would've been flagged:

```
RU
CN
```

```
AU
HK
```

And RU would've been flagged twice, for it not only is far above the threshold but is also a new entry in the list.

The IP addresses that connected to the Russian domains were:

```
192.168.105.76
192.168.105.159
```

Ports used for communications

Ports used

Unexpected ports being in use may be a clear sign of infection, as malware often makes use of uncommon ports, to avoid using ports that are already in use on the machine, risking detection or malfunction.
Different distributions in port usage may also be an anomaly sign.

Results

Non-Anomalous data:

```
Ports Used for Communications:
443    853677
53     115248

Ports Used for Communications (%):
443    88.105581
53     11.894419
```

Test data:

```
Ports Used for Communications:
443    845186
53     182634

Ports Used for Communications (%):
443    82.230935
53     17.769065
```

Detection: Any port that's not present in the non-anomalous behaviour analysis.

With this criteria applied, nothing would've been flagged.

To be noted that a notable difference can be seen in the % values for each port. To pinpoint the anomalous traffic, a machine-by-machine analysis and criteria are needed, but are better suited for analyzing volume of data (bytes) by port instead of raw packet number.

Volume of Data, per Port

Unexpectedly high volume of data being transferred through a port not usually used may be suspicious, as it is likely to be an exfiltration attempt, or if receiving, serving as a hop in the chain for exfiltrated data.

Results

Non-Anomalous Data:

```
Volume of Data by Port:
        up_bytes   down_bytes
port
53      23076267     52898880
443   9727702861  89806260682
Volume of Data by Port (%):
      up_bytes  down_bytes
port
53    0.023167    0.053106
443   9.765795   90.157932
```

Test Data:

```
Volume of Data by Port:
         up_bytes   down_bytes
port
53       36576570     84087780
443   16797374743  88671248144

Volume of Data by Port (%):
       up_bytes  down_bytes
port
53     0.034640    0.079637
443   15.908219   83.977504
```

Variation in (absolute)% of bytes per machine:

```
                     up_bytes_variation  down_bytes_variation
src_ip          port
192.168.105.141 53             22.839858              6.818845
192.168.105.106 53             10.747322              2.847368
192.168.105.168 443             1.186653              0.258434
```

```
192.168.105.21  53              0.852324              0.201611
192.168.105.89  443             0.271583              0.067218
192.168.105.129 53              0.269499              0.066226
192.168.105.188 443             0.247848              0.009944
192.168.105.20  443             0.236574              0.006924
192.168.105.164 53              0.198914              0.030140
192.168.105.118 443             0.167755              0.002987
192.168.105.182 443             0.145731              0.011068
192.168.105.183 443             0.145504              0.000636
192.168.105.91  53              0.079763              0.019994
192.168.105.92  53              0.073666              0.017211
192.168.105.59  53              0.066781              0.018648
192.168.105.116 53              0.065654              0.020535
192.168.105.153 443             0.057249              0.019355
192.168.105.81  443             0.056784              0.012833
...
```

Detection: Threshold in % of total data being surpassed as a preliminary alert, after which the anomalous machines can be identified by monitoring data flows.

For a threshold of an 1% variation, we get the following results flagged:

```
                     up_bytes_variation  down_bytes_variation
src_ip          port
192.168.105.106 53             10.747322              2.847368
192.168.105.141 53             22.839858              6.818845
192.168.105.168 443             1.186653              0.258434
```

Volume of Data up/downloaded, by source IP

Unexpectedly high volume of data being transferred might be an data exfiltration , attempt or a dropper downloaded unwanted malware, as with the previous metric.

Results

Non-Anomalous Data:

```
Total Volume of Data: 99609938690
Volume of Data per Machine:
                  up_bytes  down_bytes
src_ip
192.168.105.100  121367563  1110984752
192.168.105.101   52724243   481635183
192.168.105.102   63662612   574736338
192.168.105.103   48623851   452207367
192.168.105.104   52250205   487287945
...                    ...         ...
192.168.105.95    51726211   463683506
192.168.105.96    49640783   469388779
192.168.105.97     7928157    81871788
192.168.105.98    57518870   554121576
192.168.105.99    19836405   175931795


Percentage of Data per Machine:
                 up_bytes  down_bytes
src_ip
192.168.105.100  0.121843    1.115335
192.168.105.101  0.052931    0.483521
192.168.105.102  0.063912    0.576987
192.168.105.103  0.048814    0.453978
192.168.105.104  0.052455    0.489196
...                   ...         ...
192.168.105.95   0.051929    0.465499
192.168.105.96   0.049835    0.471227
192.168.105.97   0.007959    0.082192
192.168.105.98   0.057744    0.556291
192.168.105.99   0.019914    0.176621
```

Test Data:

```
Total Volume of Data: 105589287237
Volume of Data per Machine:
                  up_bytes  down_bytes
src_ip
192.168.105.100   40065077   370138927
192.168.105.101   67735185   607768710
192.168.105.102   86681126   780124830
192.168.105.103   79774023   743232422
192.168.105.104   16268683   155817779
...                    ...         ...
192.168.105.95   128102638  1206773962
192.168.105.96    29075503   259457396
192.168.105.97    56689987   520181351
192.168.105.98    49657884   467606040
192.168.105.99    47233951   440373341


Percentage of Data per Machine:
                 up_bytes  down_bytes
src_ip
192.168.105.100  0.037944    0.350546
192.168.105.101  0.064150    0.575597
192.168.105.102  0.082093    0.738830
192.168.105.103  0.075551    0.703890
192.168.105.104  0.015408    0.147570
...                   ...         ...
192.168.105.95   0.121322    1.142894
192.168.105.96   0.027536    0.245723
192.168.105.97   0.053689    0.492646
192.168.105.98   0.047029    0.442854
192.168.105.99   0.044734    0.417063
```

Variation

```
src_ip           up_bytes_variation  down_bytes_variation
192.168.105.20             5.171601             -0.047975
192.168.105.188            1.692113             -0.505467
192.168.105.183            0.179811              0.463227
192.168.105.118            0.178077              0.408568
192.168.105.95             0.076676              0.746000
192.168.105.135            0.075159              0.689556
                                                       ...
192.168.105.100           -0.081621             -0.743747
```

```
192.168.105.75            -0.083654             -0.768529
192.168.105.55            -0.087691             -0.814792
192.168.105.203           -0.090302             -0.825784
192.168.105.181           -0.092976             -0.844899
192.168.105.206           -0.112453             -1.012440
```

Detection: Threshold in % being surpassed, for either up or download.

For a threshold of (absolute) 1%, the following would be flagged:

```
                up_bytes_variation  down_bytes_variation
src_ip
192.168.105.188            1.692113             -0.505467
192.168.105.20             5.171601             -0.047975
192.168.105.206           -0.112453              -1.01244
```

Protocol used for communication.

Unexpected protocols being used may be a sign of infection, as an intruder in the network may make use of protocols such as Telnet for communications with a reverse shell, which would not show up as http/tcp traffic.

Results

Non-Anomalous Data:

```
Protocol Used for Communications:
tcp    851395
udp    117530

Protocol Used for Communications (%):
tcp    87.870062
udp    12.129938
```

Test Data:

```
Protocol Used for Communications:
tcp    842895
udp    184925
Name: proto, dtype: int64
Protocol Used for Communications (%):
tcp    82.008036
udp    17.991964
Name: proto, dtype: float64
```

Variation:
(Between distribution for udp/tcp for each IP address)
(Since UDP+TCP=100%, the increases are mirrored (negative) on the other port that's not present below.)

```
                      up_bytes_variation  down_bytes_variation
src_ip          proto
192.168.105.141 udp             22.427070              4.383402
192.168.105.106 udp             10.908184              4.157021
192.168.105.164 udp              2.134167             11.628416
192.168.105.40  tcp              1.012022              8.185145
192.168.105.168 udp             -0.264823              7.498026
192.168.105.165 tcp              0.776530              7.145632
192.168.105.81  tcp              0.728540              6.830527
```

```
192.168.105.97  tcp              0.715462              6.678861
192.168.105.59  tcp              0.822587              6.647362
192.168.105.153 udp              0.630451              6.189669
192.168.105.93  udp              0.706733              5.881039
192.168.105.57  udp              0.700359              5.728980
192.168.105.185 udp              0.594755              5.626358
192.168.105.41  tcp              0.669780              5.499036
```

Detection: Any protocol that's not used in the non-anomalous behaviour analysis, and a threshold on the variation of the balance between the protocols.

For a threshold of 7%, the following rows are filtered:

```
                     up_bytes_variation  down_bytes_variation
src_ip          proto
192.168.105.106 udp             10.908184              4.157021
192.168.105.141 udp             22.427070              4.383402
192.168.105.164 udp              2.134167             11.628416
192.168.105.165 tcp              0.776530              7.145632
192.168.105.168 udp             -0.264823              7.498026
192.168.105.40  tcp              1.012022              8.185145
```

Except the 10%> variations, these values seem to fluctuate a lot and as such wouldn't be enough by themselves to flag a machine.

Visiting domains that have uncommon TLDs

While not guaranteed to be malicious, it may be a sign the user is being phished, or that the user is visiting a malicious website, while not needing an exhaustive whitelist for the domains themselves.

Results

Non Anomalous Data:

```
[('com', 15), ('net', 16), ('org', 1), ('pt', 8)] Total: 40
```

Test Data:

```
[('com', 9), ('net', 15), ('org', 1), ('pt', 7)] Total: 32
```

Variation:
The TLDs present in the resolved domains are the same in both datasets, with a slightly different distribution. However, the number of resolved domains is low enough for this data not to be significant enough to make decisions with base on.

Detection: Same as with geolocation-based analysis, plus a blacklist of uncommon TLDs.

Conclusions

Based on the completed analysis, these were the devices flagged by the set rules:

Device	# Rule Hits
192.168.105.106	5
192.168.105.141	5
192.168.105.168	2
192.168.105.53	2
192.168.105.188	1
192.168.105.443	1
192.168.105.20	1
192.168.105.206	1
192.168.105.164	1
192.168.105.165	1
192.168.105.168	1
192.168.105.40	1
192.168.105.76	1
192.168.105.159	1

Taking these results into account, it is safe to assume that .106 and .141 are machines that are certainly behaving in anomalous ways, and with .168 and .53 being *possibly* anomalous.