Botnet Detection Algorithm

The objective of this algorithm is to identifying highly periodic edges from network logs at scale which are likely to be part of the beginning of a botnet attack. The main strategy of this algorithm is detection of C2 traffic which is a precursor to a botnet attack. This behavior is needed for the bot to: - 1) update their data, 2) receive commands, 3) send keep-alive messages. The behavior is observed when looking at the transport port of the of the bot for its C2 communication.

Method

Compute periodogram which identifies the peak of the frequency domain of the discrete time series traffic signal. After peak is located, Walker's large sample test is applied to confirm that the peak is significant enough compared to the rest of the periodogram's ordinates. We compute the peak and confirm with Walker's test for different discrete time intervals. The peak of the frequency domain is obtained by computing the power spectral density of the network traffic which can be estimated using the Fourier transform of the autocorrelation function or a periodogram. The periodogram of a time sequence gives its power at different frequencies.

We used LANL dataset to validate this algorithm. A typical day's network traffic log looks like this:

- Duration 24 hours
- Hosts = 8906
- 73,784 edges
- No. of events = 9 Million

In order to compute the periodogram there is a need to reduce events as the computation is costly – it costs about 2 secs to compute the periodogram of one edge. We developed two segmentation techniques which work well in this regard.

Segmentation 1: identify unidirectional edges

- Firewalls will only allow traffic going out and not coming in especially if its dubious traffic that might be coming from a malicious site or unknown site or poor reputation sites etc.
- Removing bidirectional edges reduces number of edges to 10,902 and number of events to 1.5 Million

Segmentation 2: remove byte/packet communications with low frequency

- Low frequency of similar byte/packet connections means there are very few 'botnet-like' connections because it points towards low repeatability.
- When communications from the same edge represents the same bytes being transmitted, this resembles botnet communications.
- By filtering highly repeatable bytes/packet within the same edge we minimize the total number of edges that are of interest.

- Isolating edges that have more than 1000 times the average number of bytes/packet for an edge reduces number of edges of interest by 99% (9 Million to 762)

Combining the two segmentation strategies, leads to a set of 107 edges that we need to compute their periodogram. Computing periodogram for 107 edges takes 150 secs on a laptop.

Compute periodogram

We compute the maximum spectral density (periodogram) at different discrete time intervals (1s – 15s). This is mainly because different time intervals affect the maximum periodogram and there is no knowing beforehand which is the ideal time interval to consider. We suspect that durations greater than one minute are less likely to be exhibited by botnets. After computing the periodogram, it is important to select only the edges that whose periodogram meets the statistical sample size requirements in order to minimize error. Also edges with high values of the periodogram computation are more likely to exhibit botnet behavior. We selected edges that had maximum periodogram readings higher than 500. This number can be adjusted accordingly depending on how many edges one can handle.

We apply walker's large test to the periodogram's maximum ordinate to determine if it is periodic or not.

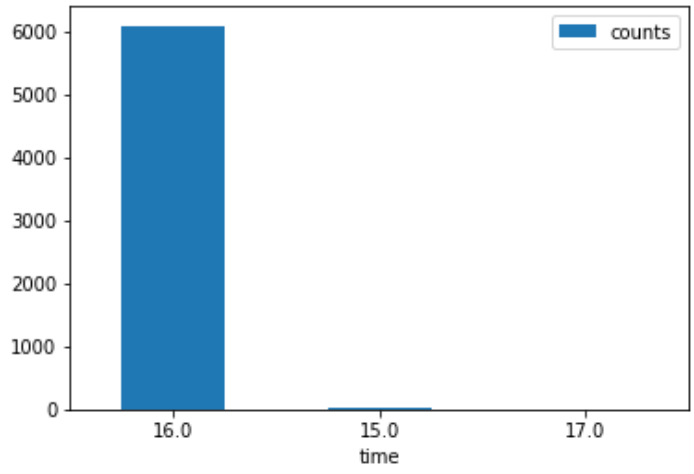
Results

These are end points that have maximum periodogram value > 500 (most likely to be periodic) in one day.

source	destination	max_pxx	samples	z_const	g_x
C1340	C787	2532.55486	9759	32.1864007	11787.5075
C1015	C15487	1471.42473	4436	30.6095271	6337.16063
C3871	C23147	1337.55053	4437	30.6099779	5477.7969
C1015	C12562	1182.15366	6099	31.2462705	4815.83527
C17693	C5074	994.418898	12201	32.6330566	4153.56633
C1015	C11114	975.757805	3254	29.989791	4030.60185
C3173	C1554	874.336046	3748	30.2724657	3577.90917
C1015	C14163	784.477269	3754	30.2756648	3479.8424
C6802	C5721	648.197084	3253	29.9891763	3477.9888
C12992	C5721	612.370687	3251	29.9879463	3329.13244
C9274	C5721	601.250239	3253	29.9891763	3320.52276
C10043	C5721	596.664785	3253	29.9891763	3305.53128
C528	C1621	592.068324	3253	29.9891763	2828.53799
C13048	C5721	585.343889	3251	29.9879463	3264.98688
C13940	C5721	576.269765	3251	29.9879463	3194.05578
C1252	C10054	513.124523	1755	28.754959	2138.42512
C8804	C5721	509.561408	3249	29.9867155	3013.15552

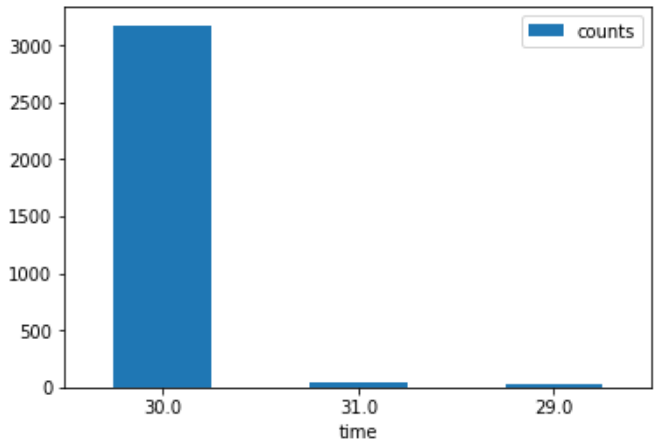
The figures below show the distribution of traffic for each end point

C1340 – C787



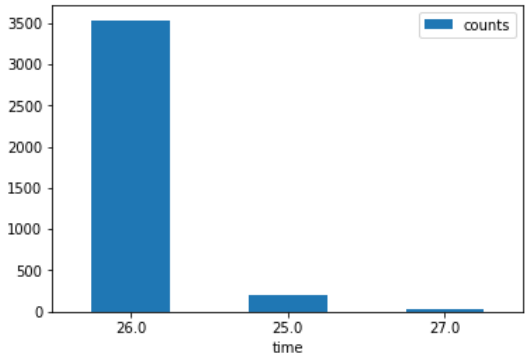
Time(sec)	counts
16	6082
15	9
17	8

C1015 – C15487



Time(sec)	counts
30	3176
31	42
29	34

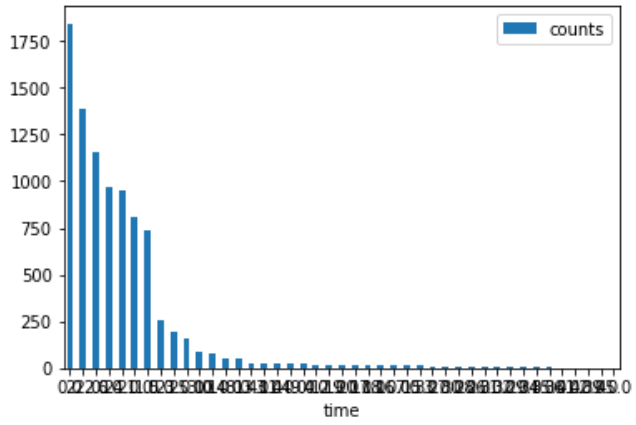
C3871 – C23147



C1015 – C12562

Time(sec)	counts
26	2534
25	203
27	96

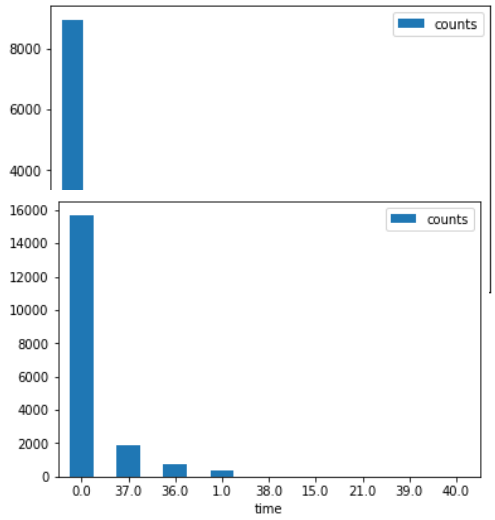
Time(sec)	counts
0	1843
22	1383
2	1153
6	968
24	949
21	812
1	735
5	258
23	196
25	158
3	90
10	74



14	54
8	54
13	29
43	28
11	27
44	23
9	22
4	20
12	18
19	17
20	17
17	16
18	14
16	14
7	13
15	12

C17693 – C5074

Time(sec)	**counts**
10	8943
11	646
9	84
12	10
13	2
14	1
16	1

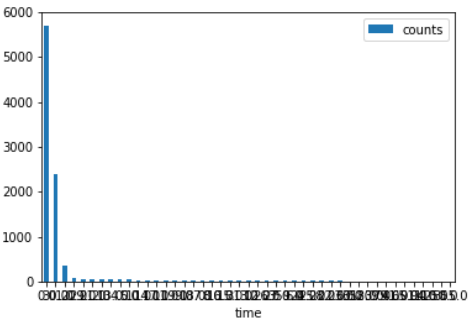


C1015 – C11114

Time(sec)	counts
0	5705
30	2382
1	351
20	70
29	54
21	48
2	45
13	40
4	36
5	36
10	34
14	33
17	32
11	31
19	31
9	31
18	30
7	30
8	30
16	30
28	1
47	1
54	1

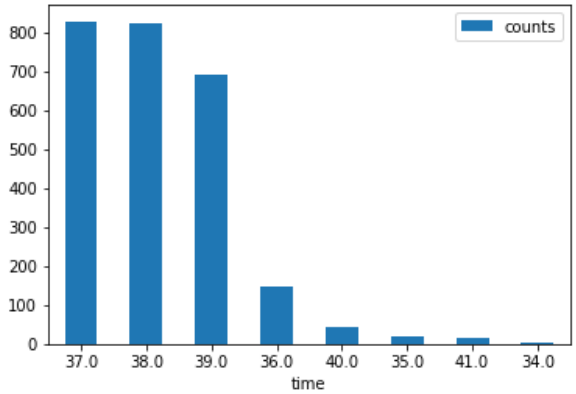
Time(sec)	counts
0	15700
37	1887
36	742
1	350
38	15
15	1
21	1
39	1
40	1

C3173 – C1554



Time(sec)	counts
37	829
38	825
39	693
36	149
40	44
35	18
41	17
34	2

C1015 – C14163



The tables below the top daily edges based on maximum periodogram values and exhibit most probable behavior of botnets.

Day 2

source	destination	max_pxx	samples	z_const	g_x
C1340	C787	1829.62881	9027	32.030461	9757.47996
C3871	C23147	1143.11517	3283	30.0075363	4682.08782
C1015	C15487	1074.47695	3611	30.1979905	4836.15077
C17693	C5074	633.526488	2683	29.6038922	2547.97366
C1015	C11114	627.044559	1810	28.8166749	2654.7163
C6177	C15348	571.692611	2483	29.4489561	2317.60996
C5474	C1970	525.4337	3611	30.1979905	2642.34366

Day 3

source	destination	max_pxx	samples	z_const	g_x
C1340	C787	2703.33478	7814	31.7418548	11534.2778
C1015	C15487	1423.83624	4687	30.7196064	6405.21553
C1015	C12562	1306.84813	6698	31.4336388	5383.05956
C3173	C1554	1230.16638	4689	30.7204596	5294.84722
C3871	C23147	1082.52852	7814	31.7418548	6090.75503
C1015	C14163	1054.58192	4262	30.529498	5116.14661
C1015	C11114	809.391918	3606	30.1952192	3530.9535
C13119	C5721	623.197984	3125	29.9088896	3343.23809
C13665	C5721	564.630901	3125	29.9088896	3153.96952
C8462	C5721	516.091875	3125	29.9088896	2981.3312
C528	C1621	500.249687	3349	30.0473446	2505.11885

Day 4

source	destination	max_pxx	samples	z_const	g_x
C1340	C787	2874.85814	14185	32.9343909	15332.0364
C3871	C23147	1738.4247	6304	31.3123896	7674.03007
C1015	C12562	1737.62473	9457	32.1235312	7555.68162
C1015	C15487	1634.70223	5675	31.102162	7357.13236
C3173	C1554	1397.31048	4728	30.7370255	5674.21338
C1015	C11114	1319.80845	4364	30.5767991	5765.55364
C1015	C14163	1186.10605	5157	30.9107309	5754.29974
C17693	C5074	989.532428	18914	33.5098255	4723.23449
C5474	C3088	652.078463	7091	31.5476736	3390.05247
C1015	C13791	537.156326	3779	30.2889398	3309.37924
C528	C1621	510.095945	4727	30.7366024	2792.43269

Day 5

source	destination	max_pxx	samples	z_const	g_x
C1340	C787	2532.55486	9759	32.1864007	11787.5075
C1015	C15487	1471.42473	4436	30.6095271	6337.16063
C3871	C23147	1337.55053	4437	30.6099779	5477.7969
C1015	C12562	1182.15366	6099	31.2462705	4815.83527
C17693	C5074	994.418898	12201	32.6330566	4153.56633
C1015	C11114	975.757805	3254	29.989791	4030.60185
C3173	C1554	874.336046	3748	30.2724657	3577.90917
C1015	C14163	784.477269	3754	30.2756648	3479.8424
C6802	C5721	648.197084	3253	29.9891763	3477.9888
C12992	C5721	612.370687	3251	29.9879463	3329.13244
C9274	C5721	601.250239	3253	29.9891763	3320.52276
C10043	C5721	596.664785	3253	29.9891763	3305.53128
C528	C1621	592.068324	3253	29.9891763	2828.53799
C13048	C5721	585.343889	3251	29.9879463	3264.98688
C13940	C5721	576.269765	3251	29.9879463	3194.05578
C1252	C10054	513.124523	1755	28.754959	2138.42512
C8804	C5721	509.561408	3249	29.9867155	3013.15552

Day 6

source	destination	max_pxx	samples	z_const	g_x
C1340	C787	1778.56294	8775	31.9738344	9485.12862
C2660	C10054	1168.60548	3534	30.1548817	4881.2981
C10	C3322	1102.62039	8775	31.9738344	4789.8216
C1015	C15487	1002.00443	3191	29.9506897	4314.91111
C3871	C23147	881.943412	2700	29.6165246	3528.44171
C1252	C10054	875.131145	4018	30.4115895	3854.6154
C6177	C15348	762.753968	5015	30.8548877	3495.48503
C5336	C5721	718.574499	7441	31.6440313	3155.92485
C1015	C11286	608.296029	2083	29.0976395	2610.25408
C1015	C14163	596.078722	2651	29.5798949	2891.41656
C1015	C20107	562.826195	2913	29.768388	2894.64937

Day 7

source	destination	max_pxx	samples	z_const	g_x
C1252	C10054	1942.98384	5585	31.0701897	8161.84041
C1340	C787	1637.95828	4042	30.4235002	6552.64377
C2660	C10054	1345.73869	5465	31.0267491	5603.67853
C1015	C15487	978.360643	3234	29.9774605	4404.3253
C3871	C23147	972.116697	2939	29.7861598	3981.96438
C5336	C5721	778.068892	10779	32.3852204	4014.18217
C6177	C15348	705.135087	3592	30.1874393	2930.13588
C10	C3322	512.175353	5390	30.9991117	2058.20271

Conclusion

The results show that as the maximum periodogram reduces, the periodic behavior on edge also reduces accordingly. Among the top periodic edges, we are able to identify an edge which is a subject of lots of redteam events which is a possible candidate of botnet behavior.