DATA602 – Advanced Programming
Michael Ippolito
4/7/2022

Final Project Proposal

Research Question

As a cybersecurity professional, I spend a portion of my day hunting for potentially compromised computers connected to my organization's network. Some of that time is spent examining traffic as it flows across the network from one IP address to another. Each flow can be characterized by a number of parameters, including source and destination port, network protocol, packets and bytes sent and received, and packet flowrate. On any network, some of that traffic will inevitably be malicious: attackers probing for vulnerabilities to exploit for various reasons (e.g., monetary profit, notoriety, or political reasons). Evaluating which of those flows contain malicious traffic on even a slightly busy network would be impossible to perform manually. Therefore, it is eminently appropriate to develop machine-learning techniques to do it programmatically.

Justification

It should go without saying that organizations have a vested interest in discovering and mitigating attacks against their networks. Further, if a machine is compromised, it also goes without saying that the machine should be identified and remediated as soon as possible after being infected. Organizations, therefore, are justified in investigating programmatic ways of monitoring their networks and taking appropriate action when evidence of malfeasance is discovered.

Data Sources

The data sets I chose to work with come from the Canadian Institute for Cybersecurity (https://www.unb.ca/cic/datasets/ids-2017.html). The data is free for the public to use and is classified as an "intrusion detection evaluation dataset," intended for audiences to train machine-learning models.

Libraries

The project will likely include the following core libraries:

- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn

- scipy

Additionally, the following helper libraries will probably also be useful:

- csv
- re
- math
- datetime
- sys
- os

Data Cleaning and Wrangling

Several cleaning and wrangling steps were performed on the raw data.

1. The data was read from eight separate CSV files and concatenated into a single data frame.
2. Field names were set to lowercase and normalized.
3. Date/time fields were converted to timestamp format.
4. The “protocol” field was converted from numeric to a categorical string.
5. Observations with NaNs or where the protocol was zero were dropped.

Exploratory Data Analysis and Summary Statistics

```
Data frame info:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2827689 entries, 0 to 692702
Data columns (total 86 columns):
 #   Column               Dtype
---  ------               -----
 0   flow_id              object
 1   srcip                object
 2   src_port             int64
 3   dstip                object
 4   dst_port             int64
 5   proto                object
 6   timestamp            datetime64[ns]
 7   flow_duration        int64
 8   tot_fwd_pkts         int64
 9   tot_bwd_pkts         int64
 10  tot_len_of_fwd_pkts  float64
 11  tot_len_of_bwd_pkts  float64
 12  fwd_pkt_len_max      float64
 13  fwd_pkt_len_min      float64
 14  fwd_pkt_len_mean     float64
 15  fwd_pkt_len_std      float64
 16  bwd_pkt_len_max      float64
 17  bwd_pkt_len_min      float64
 18  bwd_pkt_len_mean     float64
 19  bwd_pkt_len_std      float64
 20  flow_bps             float64
```

```
21  flow_pps             float64
22  flow_iat_mean        float64
23  flow_iat_std         float64
24  flow_iat_max         float64
25  flow_iat_min         float64
26  fwd_iat_tot          float64
27  fwd_iat_mean         float64
28  fwd_iat_std          float64
29  fwd_iat_max          float64
30  fwd_iat_min          float64
31  bwd_iat_tot          float64
32  bwd_iat_mean         float64
33  bwd_iat_std          float64
34  bwd_iat_max          float64
35  bwd_iat_min          float64
36  fwd_psh_flags        int64
37  bwd_psh_flags        int64
38  fwd_urg_flags        int64
39  bwd_urg_flags        int64
40  fwd_hdr_len          int64
41  bwd_hdr_len          int64
42  fwd_pps              float64
43  bwd_pps              float64
44  min_pkt_len          float64
45  max_pkt_len          float64
46  pkt_len_mean         float64
47  pkt_len_std          float64
48  pkt_len_var          float64
49  fin_flag_ct          int64
50  syn_flag_ct          int64
51  rst_flag_ct          int64
52  psh_flag_ct          int64
53  ack_flag_ct          int64
54  urg_flag_ct          int64
55  cwe_flag_ct          int64
56  ece_flag_ct          int64
57  down_up_ratio        float64
58  avg_pkt_sz           float64
59  avg_fwd_seg_sz       float64
60  avg_bwd_seg_sz       float64
61  fwd_hdr_len_1        int64
62  fwd_avg_bytes_bulk   int64
63  fwd_avg_pkts_bulk    int64
64  fwd_avg_bulk_rate    int64
65  bwd_avg_bytes_bulk   int64
66  bwd_avg_pkts_bulk    int64
67  bwd_avg_bulk_rate    int64
68  subflow_fwd_pkts     int64
69  subflow_fwd_bytes    int64
70  subflow_bwd_pkts     int64
71  subflow_bwd_bytes    int64
72  init_win_bytes_fwd   int64
73  init_win_bytes_bwd   int64
74  act_data_pkt_fwd     int64
75  min_seg_sz_fwd       int64
76  active_mean          float64
77  active_std           float64
```

```
 78  active_max           float64
 79  active_min           float64
 80  idle_mean            float64
 81  idle_std             float64
 82  idle_max             float64
 83  idle_min             float64
 84  label                object
 85  fn                   object
dtypes: datetime64[ns](1), float64(45), int64(34), object(6)
memory usage: 1.8+ GB
None

Proto categories:
['tcp' 'udp']

Label categories:
['BENIGN' 'DDoS' 'PortScan' 'Bot' 'Infiltration'
 'Web Attack \x96 Brute Force' 'Web Attack \x96 XSS'
 'Web Attack \x96 Sql Injection' 'FTP-Patator' 'SSH-Patator'
 'DoS slowloris' 'DoS Slowhttptest' 'DoS Hulk' 'DoS GoldenEye'
 'Heartbleed']

Summary stats:
                                        flow_id       srcip      src_port  \
count                                   2827689     2827689  2.827689e+06
unique                                  1084943       16993           NaN
top     192.168.10.255-192.168.10.3-137-137-17  172.16.0.1           NaN
freq                                       523      558279           NaN
mean                                       NaN         NaN  4.115386e+04
min                                        NaN         NaN  1.000000e+00
25%                                        NaN         NaN  3.281400e+04
50%                                        NaN         NaN  5.095400e+04
75%                                        NaN         NaN  5.842200e+04
max                                        NaN         NaN  6.553500e+04
std                                        NaN         NaN  2.228052e+04

                dstip      dst_port    proto                       timestamp  \
count         2827689  2.827689e+06  2827689                         2827689
unique          19041           NaN        2                             NaN
top     192.168.10.3           NaN      tcp                             NaN
freq          685169           NaN  1828196                             NaN
mean             NaN  8.076090e+03      NaN  2017-05-11 07:46:08.270103296
min              NaN  1.000000e+00      NaN            2017-03-07 01:00:01
25%              NaN  5.300000e+01      NaN            2017-04-07 04:25:00
50%              NaN  8.000000e+01      NaN            2017-05-07 10:48:00
75%              NaN  4.430000e+02      NaN            2017-06-07 12:44:00
max              NaN  6.553500e+04      NaN            2017-07-07 12:59:00
std              NaN  1.828785e+04      NaN                            NaN

        flow_duration  tot_fwd_pkts  tot_bwd_pkts  tot_len_of_fwd_pkts  \
count    2.827689e+06  2.827689e+06  2.827689e+06         2.827689e+06
unique            NaN           NaN           NaN                  NaN
top               NaN           NaN           NaN                  NaN
freq              NaN           NaN           NaN                  NaN
mean     1.475058e+07  9.307030e+00  1.040477e+01         5.498957e+02
min     -1.300000e+01  1.000000e+00  0.000000e+00         0.000000e+00
25%      1.550000e+02  2.000000e+00  1.000000e+00         1.200000e+01
```

50%	3.131700e+04	2.000000e+00	2.000000e+00	6.200000e+01
75%	3.178160e+06	5.000000e+00	4.000000e+00	1.880000e+02
max	1.200000e+08	2.197590e+05	2.919220e+05	1.290000e+07
std	3.360454e+07	7.500715e+02	9.979267e+02	9.998968e+03

	tot_len_of_bwd_pkts	fwd_pkt_len_max	fwd_pkt_len_min	\
count	2.827689e+06	2.827689e+06	2.827689e+06	
unique	NaN	NaN	NaN	
top	NaN	NaN	NaN	
freq	NaN	NaN	NaN	
mean	1.618010e+04	2.078242e+02	1.873387e+01	
min	0.000000e+00	0.000000e+00	0.000000e+00	
25%	2.000000e+00	6.000000e+00	0.000000e+00	
50%	1.230000e+02	3.700000e+01	2.000000e+00	
75%	4.840000e+02	8.100000e+01	3.600000e+01	
max	6.554530e+08	2.482000e+04	2.325000e+03	
std	2.264310e+06	7.175396e+02	6.036878e+01	

	fwd_pkt_len_mean	fwd_pkt_len_std	bwd_pkt_len_max	bwd_pkt_len_min
\				
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN
mean	5.826480e+01	6.898456e+01	8.717900e+02	4.109391e+01
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	6.000000e+00	0.000000e+00	2.000000e+00	0.000000e+00
50%	3.400000e+01	0.000000e+00	7.900000e+01	0.000000e+00
75%	5.000000e+01	2.616295e+01	2.820000e+02	7.700000e+01
max	5.940857e+03	7.125597e+03	1.953000e+04	2.896000e+03
std	1.861818e+02	2.813298e+02	1.947208e+03	6.888656e+01

	bwd_pkt_len_mean	bwd_pkt_len_std	flow_bps	flow_pps	\
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06	
unique	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	
mean	3.062797e+02	3.356879e+02	inf	inf	
min	0.000000e+00	0.000000e+00	-2.610000e+08	-2.000000e+06	
25%	2.000000e+00	0.000000e+00	1.194046e+02	3.474790e+00	
50%	7.200000e+01	0.000000e+00	4.605455e+03	1.105767e+02	
75%	1.810000e+02	7.850000e+01	1.666667e+05	2.325581e+04	
max	5.800500e+03	8.194660e+03	inf	inf	
std	6.055000e+02	8.400742e+02	NaN	NaN	

	flow_iat_mean	flow_iat_std	flow_iat_max	flow_iat_min	fwd_iat_tot
\					
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	1.298953e+06	2.920525e+06	9.183669e+06	1.625012e+05	1.444756e+07
min	-1.300000e+01	0.000000e+00	-1.300000e+01	-1.400000e+01	0.000000e+00
25%	6.366667e+01	0.000000e+00	1.240000e+02	3.000000e+00	0.000000e+00
50%	1.144700e+04	1.371787e+02	3.086600e+04	4.000000e+00	4.300000e+01
75%	3.359287e+05	6.860236e+05	2.398884e+06	6.400000e+01	1.225072e+06
max	1.200000e+08	8.480026e+07	1.200000e+08	1.200000e+08	1.200000e+08

std	4.509397e+06	8.048708e+06	2.446954e+07	2.951645e+06	3.352607e+07

	fwd_iat_mean	fwd_iat_std	fwd_iat_max	fwd_iat_min	bwd_iat_tot
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	2.612070e+06	3.268587e+06	9.043985e+06	1.022903e+06	9.904175e+06
min	0.000000e+00	0.000000e+00	0.000000e+00	-1.200000e+01	0.000000e+00
25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	2.550000e+01	0.000000e+00	4.200000e+01	3.000000e+00	3.000000e+00
75%	2.041860e+05	6.533096e+04	9.114870e+05	4.800000e+01	9.927400e+04
max	1.200000e+08	8.460293e+07	1.200000e+08	1.200000e+08	1.200000e+08
std	9.530091e+06	9.642980e+06	2.453921e+07	8.595742e+06	2.875008e+07

	bwd_iat_mean	bwd_iat_std	bwd_iat_max	bwd_iat_min	fwd_psh_flags
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	1.807680e+06	1.487498e+06	4.689560e+06	9.682957e+05	4.649663e-02
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	3.000000e+00	0.000000e+00	3.000000e+00	1.000000e+00	0.000000e+00
75%	1.836575e+04	1.584338e+04	6.062500e+04	4.500000e+01	0.000000e+00
max	1.200000e+08	8.441801e+07	1.200000e+08	1.200000e+08	1.000000e+00
std	8.891783e+06	6.281561e+06	1.716932e+07	8.313407e+06	2.105581e-01

	bwd_psh_flags	fwd_urg_flags	bwd_urg_flags	fwd_hdr_len
count	2827689.0	2.827689e+06	2827689.0	2.827689e+06
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN
mean	0.0	1.113984e-04	0.0	-2.602550e+04
min	0.0	0.000000e+00	0.0	-3.221223e+10
25%	0.0	0.000000e+00	0.0	4.000000e+01
50%	0.0	0.000000e+00	0.0	6.400000e+01
75%	0.0	0.000000e+00	0.0	1.200000e+02
max	0.0	1.000000e+00	0.0	4.644908e+06
std	0.0	1.055396e-02	0.0	2.106422e+07

	bwd_hdr_len	fwd_pps	bwd_pps	min_pkt_len	max_pkt_len
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	-2.275732e+03	6.390901e+04	7.002742e+03	1.645225e+01	9.514289e+02
min	-1.073741e+09	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.000000e+01	1.766016e+00	1.262489e-01	0.000000e+00	6.000000e+00
50%	4.000000e+01	6.155253e+01	1.987736e+01	2.000000e+00	8.700000e+01
75%	1.040000e+02	1.204819e+04	7.380074e+03	3.600000e+01	5.320000e+02
max	5.838440e+06	3.000000e+06	2.000000e+06	1.448000e+03	2.482000e+04
std	1.452993e+06	2.476127e+05	3.817160e+04	2.524556e+01	2.029083e+03

```
        pkt_len_mean   pkt_len_std   pkt_len_var   fin_flag_ct   syn_flag_ct
\
count   2.827689e+06  2.827689e+06  2.827689e+06  2.827689e+06  2.827689e+06
unique           NaN           NaN           NaN           NaN           NaN
top              NaN           NaN           NaN           NaN           NaN
freq             NaN           NaN           NaN           NaN           NaN
mean    1.721301e+02  2.952941e+02  4.866799e+05  3.540099e-02  4.649663e-02
min     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%     6.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%     5.733333e+01  2.629068e+01  6.912000e+02  0.000000e+00  0.000000e+00
75%     1.198095e+02  1.752712e+02  3.072000e+04  0.000000e+00  0.000000e+00
max     3.337143e+03  4.731522e+03  2.240000e+07  1.000000e+00  1.000000e+00
std     3.056041e+02  6.320667e+02  1.648302e+06  1.847912e-01  2.105581e-01

         rst_flag_ct   psh_flag_ct   ack_flag_ct   urg_flag_ct   cwe_flag_ct
\
count   2.827689e+06  2.827689e+06  2.827689e+06  2.827689e+06  2.827689e+06
unique           NaN           NaN           NaN           NaN           NaN
top              NaN           NaN           NaN           NaN           NaN
freq             NaN           NaN           NaN           NaN           NaN
mean    2.426009e-04  2.983924e-01  3.157214e-01  9.489728e-02  1.113984e-04
min     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%     0.000000e+00  1.000000e+00  1.000000e+00  0.000000e+00  0.000000e+00
max     1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
std     1.557377e-02  4.575527e-01  4.648026e-01  2.930731e-01  1.055396e-02

         ece_flag_ct  down_up_ratio    avg_pkt_sz  avg_fwd_seg_sz  \
count   2.827689e+06   2.827689e+06  2.827689e+06    2.827689e+06
unique           NaN            NaN           NaN             NaN
top              NaN            NaN           NaN             NaN
freq             NaN            NaN           NaN             NaN
mean    2.436619e-04   6.841410e-01  1.921910e+02    5.826480e+01
min     0.000000e+00   0.000000e+00  0.000000e+00    0.000000e+00
25%     0.000000e+00   0.000000e+00  7.750000e+00    6.000000e+00
50%     0.000000e+00   1.000000e+00  7.225000e+01    3.400000e+01
75%     0.000000e+00   1.000000e+00  1.495000e+02    5.000000e+01
max     1.000000e+00   1.560000e+02  3.893333e+03    5.940857e+03
std     1.560777e-02   6.805038e-01  3.319795e+02    1.861818e+02

        avg_bwd_seg_sz  fwd_hdr_len_1  fwd_avg_bytes_bulk  fwd_avg_pkts_bulk
\
count     2.827689e+06   2.827689e+06           2827689.0          2827689.0
unique             NaN            NaN                 NaN                NaN
top                NaN            NaN                 NaN                NaN
freq               NaN            NaN                 NaN                NaN
mean      3.062797e+02  -2.602550e+04                 0.0                0.0
min       0.000000e+00  -3.221223e+10                 0.0                0.0
25%       2.000000e+00   4.000000e+01                 0.0                0.0
50%       7.200000e+01   6.400000e+01                 0.0                0.0
75%       1.810000e+02   1.200000e+02                 0.0                0.0
max       5.800500e+03   4.644908e+06                 0.0                0.0
std       6.055000e+02   2.106422e+07                 0.0                0.0

        fwd_avg_bulk_rate  bwd_avg_bytes_bulk  bwd_avg_pkts_bulk  \
count           2827689.0           2827689.0          2827689.0
```

unique	NaN	NaN	NaN
top	NaN	NaN	NaN
freq	NaN	NaN	NaN
mean	0.0	0.0	0.0
min	0.0	0.0	0.0
25%	0.0	0.0	0.0
50%	0.0	0.0	0.0
75%	0.0	0.0	0.0
max	0.0	0.0	0.0
std	0.0	0.0	0.0

	bwd_avg_bulk_rate	subflow_fwd_pkts	subflow_fwd_bytes
count	2827689.0	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN
top	NaN	NaN	NaN
freq	NaN	NaN	NaN
mean	0.0	9.307030e+00	5.498852e+02
min	0.0	1.000000e+00	0.000000e+00
25%	0.0	2.000000e+00	1.200000e+01
50%	0.0	2.000000e+00	6.200000e+01
75%	0.0	5.000000e+00	1.880000e+02
max	0.0	2.197590e+05	1.287034e+07
std	0.0	7.500715e+02	9.985442e+03

	subflow_bwd_pkts	subflow_bwd_bytes	init_win_bytes_fwd
count	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN
top	NaN	NaN	NaN
freq	NaN	NaN	NaN
mean	1.040477e+01	1.617976e+04	6.995787e+03
min	0.000000e+00	0.000000e+00	-1.000000e+00
25%	1.000000e+00	2.000000e+00	-1.000000e+00
50%	2.000000e+00	1.230000e+02	2.510000e+02
75%	4.000000e+00	4.840000e+02	8.192000e+03
max	2.919220e+05	6.554530e+08	6.553500e+04
std	9.979267e+02	2.264279e+06	1.434294e+04

	init_win_bytes_bwd	act_data_pkt_fwd	min_seg_sz_fwd	active_mean
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN
mean	1.990839e+03	5.424070e+00	-2.744664e+03	7.914053e+04
min	-1.000000e+00	0.000000e+00	-5.368707e+08	0.000000e+00
25%	-1.000000e+00	0.000000e+00	2.000000e+01	0.000000e+00
50%	-1.000000e+00	1.000000e+00	2.400000e+01	0.000000e+00
75%	2.350000e+02	2.000000e+00	3.200000e+01	0.000000e+00
max	6.553500e+04	2.135570e+05	1.380000e+02	1.100000e+08
std	8.459096e+03	6.367692e+02	1.085575e+06	6.347333e+05

	active_std	active_max	active_min	idle_mean	idle_std
count	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06	2.827689e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	3.865095e+04	1.465260e+05	5.817292e+04	8.319377e+06	5.025865e+05

```
min     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
max     7.420000e+07  1.100000e+08  1.100000e+08  1.200000e+08  7.690000e+07
std     3.687977e+05  9.597198e+05  5.766073e+05  2.364058e+07  4.604260e+06
```

```
            idle_max      idle_min    label  \
count   2.827689e+06  2.827689e+06  2827689
unique           NaN           NaN       15
top              NaN           NaN   BENIGN
freq             NaN           NaN  2270998
mean    8.696449e+06  7.924665e+06      NaN
min     0.000000e+00  0.000000e+00      NaN
25%     0.000000e+00  0.000000e+00      NaN
50%     0.000000e+00  0.000000e+00      NaN
75%     0.000000e+00  0.000000e+00      NaN
max     1.200000e+08  1.200000e+08      NaN
std     2.437683e+07  2.337389e+07      NaN
```

```
                                          fn
count                                2827689
unique                                     8
top     Wednesday-workingHours.pcap_ISCX.csv
freq                                  691365
mean                                     NaN
min                                      NaN
25%                                      NaN
50%                                      NaN
75%                                      NaN
max                                      NaN
std                                      NaN
```