Real-time Network Intrusion Detection with SketchFlow Sampler

Jumabek Alikhanov [*INHA], *DaeHong Min* [*INHA], *DaeHun Nyang* [*INHA]
INHA University

1. Abstract

- This work proposes using SketchFlow – flow level packet sampler together with Machine Learning classifier for real-time network intrusion detection system (NIDS).
- Our contributions are the following: 1) we provide real-time NIDS by reducing memory requirement in network switch for collecting flow features. Ability of Sketchflow for reducing data size by sampling and filtering out mice flows gives us the ability to detect intrusion in real-time; 2) through the experiments we show effectiveness of SketchFlow sampler for intrusion Detection. We also show feature estimation accuracy on Classification task.

2. Introduction

- **Motivation**

  Number of internet connected devices increasing daily. This brings heavy load of traffic to networks which will increase chance of intrusion. There are two obstacles in providing real-time NIDS for heavy traffic networks:

  1)low memory restrictions while collecting flow statistics on expensive network switch memories;

  2)large number of packets where analyzing them in real-time is not possible

- **Contribution**

  We provide single solution for the two key problems in building real-time NIDS by utilizing SketchFlow – a flow level packet sampler:

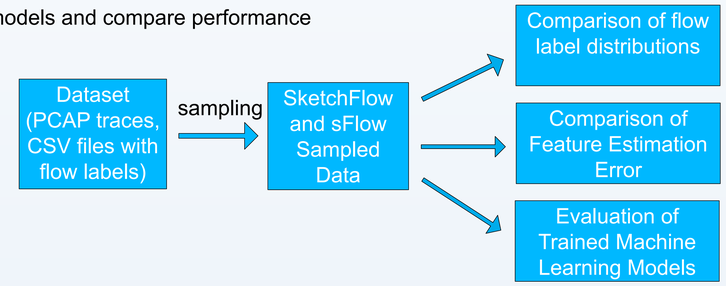
  1)we collect flow statistics with significantly small network switch memory by using SketchFlow sampler to filter out mice flows;

  2) we analyze network traffic in real time by processing only the portion of the traffic that is diligently sampled by SketchFlow;

3. Workflow

- Objective of this work is to show applicability of SketchFlow (flow-level sampler) for real-time NID by comparing it performance against sFlow (random sampler) by:

  1) Sample data
  2) Analyze sampled flow labels
  3) Estimate flow features and compare estimation error
  4) Train ML models and compare performance



4. Experimental Results

- SketchFlows ability to sample out Benign Flows is shown in Table 1. Similarly, Table 2 demonstrates effect of sampling methods on preserving the attacks flows. Table 3 shows SketchFlow works especially better when flow lengths are larger.

Table 1. Sampled Benign Flows

Num. benign flows Sketchflow	Num. benign flows sFlow	Difference	Diffrence in %
46382	57329	10907	19%

Table 2. Sampled Attack Flows

Flow Label	Num. flows sketchflow	Num. flows sFlow	Difference	Difference in %
Bot	1	16	-15	-94%
DDoS	5780	5624	156	3%
DoS Hulk	4331	4252	79	2%
FTP-Patator	547	584	-37	-6%
Infiltration	28	22	6	27%
PortScan	11	12	-1	-8%
SSH-Patator	1386	1298	88	7%
Web Attack - Brute Force	46	46	0	0%
Web Attack – XSS	9	9	0	0%
Total Attacks	12139	11863	276	2%

Table 4. Sketchflow Confusion Matrix

Ground Truth \ Prediction	Benign	Attack
Benign	11141	324
Attack	942	1430

Table 5. Sflow Confusion Matrices

Ground Truth \ Prediction	Benign	Attack
Benign	11141	324
Attack	942	1430

Table 6. Intrusion Detection Accuracy

Methods\Metric	Precision	Recall	F1 Score
SketchFlow	0.83	0.75	0.79
Sflow	0.82	0.60	0.69

***Table 3. Comparison of* Estimation Error(%)**

Feature Name \ Error %	Min. flow length =10	Min. flow length = 50	Min. flow length = 100
Tot Fwd Pkts - Sketchflow Est	16.88	7.38	5.37
Tot Fwd Pkts - SFlow Est	22.68	10.63	7.79
Tot Bwd Pkts - Sketchflow Est	18.02	6.84	5.18
Tot Bwd Pkts - SFlow Est	21.76	9.59	7.09
TotLen Fwd Pkts - Sketchflow Est	80.39	47.64	43.41
TotLen Fwd Pkts - SFlow Est	85.79	49.8	45.58
TotLen Bwd Pkts - Sketchflow Est	13.5	6.94	5.5
TotLen Bwd Pkts - SFlow Est	18.06	9.69	7.06
Fwd Pkt Len Mean - Sketchflow Est	98.54	50.06	44.64
Fwd Pkt Len Mean - SFlow Est	97.5	49.3	45.13
Fwd Pkt Len Std - Sketchflow Est	87	51.19	46
Fwd Pkt Len Std - SFlow Est	82.56	50.21	44.86
Bwd Pkt Len Mean - Sketchflow Est	32.33	8.99	5.79
Bwd Pkt Len Mean - SFlow Est	37.19	9.75	6.04
Bwd Pkt Len Std - Sketchflow Est	46.56	15.34	11.17
Bwd Pkt Len Std - SFlow Est	51.37	16.18	12.65
Fwd Header Len - Sketchflow Est	17.17	7.41	5.34
Fwd Header Len - SFlow Est	22.81	10.64	7.75
Bwd Header Len - Sketchflow Est	19.31	7.21	5.38
Bwd Header Len - SFlow Est	23.4	10.06	7.43
Pkt Len Mean - Sketchflow Est	33.33	10.54	7.1
Pkt Len Mean - SFlow Est	39.48	14.48	9.65
Pkt Len Std - Sketchflow Est	35.55	9.23	6.21
Pkt Len Std - SFlow Est	38.68	9.79	6.48
Pkt Len Var - Sketchflow Est	48.47	18.32	12.23
Pkt Len Var - SFlow Est	50.63	19.41	12.48
Down/Up Ratio - Sketchflow Est	137.39	24.65	17.52
Down/Up Ratio - SFlow Est	128.67	41.88	29.89
Pkt Size Avg - Sketchflow Est	35.69	10.26	6.93
Pkt Size Avg - SFlow Est	41.48	14.17	9.53
Fwd Seg Size Avg - Sketchflow Est	98.54	50.06	44.64
Fwd Seg Size Avg - SFlow Est	97.5	49.3	45.13
Bwd Seg Size Avg - Sketchflow Est	32.33	8.99	5.79
Bwd Seg Size Avg - SFlow Est	37.19	9.75	6.04
Fwd Act Data Pkts - Sketchflow Est	55.3	35.85	28.59
Fwd Act Data Pkts - SFlow Est	63.35	39.23	32.7
Fwd Seg Size Min - Sketchflow Est	0	0	0
Fwd Seg Size Min - SFlow Est	0	0	0

5. Analysis

- **Keeping Switch Memory Small**

  As shown in Table 1, Thanks to flow-level sampling nature, Sketchflow filters out 19% more small benign flows compared to sFlow. This means statistics of small flows (smaller than SR) will not be collected in switch memory.

- **Collecting flow statistics real-time with accuracy**

  Any type of sampling method can reduce data size therefore providing real-time processing. However, for intrusion detection we also need feature estimation to be accurate as possible. As Table 3 shows Sketchflow estimates features more accurately than sFlow, especially in larger flows (refer to last column – minimum flow length = 100)

  Intrusion Detection Accuracy in Table 6 also shows significant gap in favor of SketchFlow between models that is trained on two different sampled data.

6. Implementation

- We use the Sampler codes of Sflow and SketchFlow that is provided in C language.
- Intrusion Detector Model
  - Sampling Rate: 10
  - Classifier: RandomForest
  - Split: Stratified K(K=5) fold
  - Library: Scikit-learn

7. Concluding Remarks

- For Intrusion Detection Sketchflow sampler is effective because it will keep switch memory low by filtering out mice flows, at the same time it accurately estimates flow statistics for Intrusion Detection
- In future, we use sequential deep learning to extract flow features directly from the sequence of packets
- We also increase the scale of our experiments