Data-Fusion for Prefix-Level Inference: A DDoS Case Study

Chris Misa (UO), Ramakrishnan Durairajan (UO), Reza Rejaie (UO),
Walter Willinger (NIKSUN, Inc.)

A direct and intuitive way to improve the efficiency of machine learning (ML)-based decision-making for network traffic monitoring tasks like detecting volumetric DDoS attacks [8, 7, 17, 16, 9, 11] is to apply inference at the prefix-level rather than at an individual source- or flow-level. Although "ambiguous" prefixes containing both attack and benign traffic may reduce accuracy, considering prefix aggregates enables exponential reduction in the number of traffic entities that must be monitored and classified. This reduction is key towards realizing ML-based inference in resource-constrained environments like programmable switch hardware [4, 5] in practice. Moreover, iterative refinement algorithms can be developed to zoom-in on ambiguous prefixes and achieve accuracy that rivals that of pure source- or flow-level approaches.

However, realising prefix-level ML solutions requires prefix-level training datasets that satisfy the following two key requirements: *(i)* Training data must reflect how features associated with attack and benign traffic "blend" to form ambiguous prefixes through aggregation (*i.e.,* when going from longer to shorter prefix lengths). *(ii)* Training and testing data must come from distinct attack scenarios to avoid cross contamination through prefix aggregation (*i.e.,* a prefix in the test set must not contain any descendent prefixes that were in the training set). Also, features such as inter-packet gap statistics are non-linear in prefix aggregation and must be computed independently based on the particular interleaving of packets in each prefix.

Unfortunately, existing datasets fail to meet both of these requirements because they often include only a small number of distinct attacker source addresses in a single prefix and often contain only a single instance or just a few instances of each type of attack. To illustrate, Table 1 shows the number of distinct prefixes at /8 - /32 levels of aggregation for several publicly available datasets commonly used in DDoS defense research studies. While these existing datasets capture realistic aggregation of either attack sources or of benign sources, they do not capture realistic blending of the two. For example, although the Booters dataset [13] provides samples of actual attacks launched by DDoS-as-a-service operations, it contains no benign traffic. More recent datasets like CIC *do* include some benign traffic, but only contain a single attack source address.

Dataset	**# Benign**				**# Attack**			
	/8	**/16**	**/24**	**/32**	**/8**	**/16**	**/24**	**/32**
CAIDA ('07) [2]	0	0	0	0	117	4 k	8.7 k	9 k
ISCX ('12) [15]	123	1590	2041	2129	6	6	9	14
Booters ('15) [13]	0	0	0	0	42	961	3 k	4.4 k
Mirai ('16) [6]	0	0	0	0	162	3.5 k	9.8 k	10 k
CIC ('17) [14]	156	922	2125	3432	1	1	1	1
CSECIC ('18) [3]	1	1	6	446	2	4	10	10
MAWILab ('19) [10]	211	30 k	3.3 m	5.3 m	0	0	0	0
CAIDA ('19) [1]	250	27 k	323 k	1.3 m	0	0	0	0
Proposed "data-fusion" method	216	30 k	3.2 m	4.8 m	179	7 k	45 k	50 k

Table 1: Number of distinct attack and benign prefixes in datasets commonly used for training and testing ML-based approaches to DDoS traffic detection.

This talk describes a practical solution to the challenging requirements of prefix-level datasets for prefix-level ML which we refer to as "data-fusion". The key idea of data-fusion is to combine two or more publicly-available datasets on a flow-level or source-address level to compensate for their limitations w.r.t. prefix-level structure. This enables producing a large number of independent attack scenarios that captures realistic blending of features under prefix aggregation and presents ambiguous prefixes to the model during training. We describe our practical experience using data-fusion to train and evaluate a novel approach to volumetric DDoS attack detection called ZAPDOS [12]. Finally, we conclude by considering several new inference techniques enabled by our data-fusion method and their potential to improve the accuracy and efficiency of ML-based inference for network traffic.

References

[1] The CAIDA UCSD anonymized Internet traces dataset - 2019. https://www.caida.org/data/monitors/passive-equinix-nyc.xml.

[2] The CAIDA UCSD "DDoS attack 2007" dataset. http://www.caida.org/data/passive/ddos-20070804_dataset.xml. Accessed: 2022.

[3] IDS 2018. https://www.unb.ca/cic/datasets/ids-2018.html. Accessed: 2023-06-22.

[4] Intel Tofino. https://www.intel.com/content/www/us/en/products/network-io/programmable-ethernet-switch.html. Accessed: 2022.

[5] Trident4 / BCM56880 series. https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56880-series. Accessed: 2022.

[6] FRGP (www.frgp.net) continuous flow dataset, IMPACT ID: USC-LANDER/Mirai-FRGP-scanning-20160908/rev10326. Provided by the USC/LANDER project (http://www.isi.edu/ant/lander)., Traces taken 2016-09-08 to 2016-10-31.

[7] Muhammad Asad, Muhammad Asim, Talha Javed, Mirza O Beg, Hasan Mujtaba, and Sohail Abbas. Deepdetect: detection of distributed denial of service attacks using deep learning. *The Computer Journal*, 63(7):983–994, 2020.

[8] Roberto Doriguzzi-Corin, Stuart Millar, Sandra Scott-Hayward, Jesus Martinez-del Rincon, and Domenico Siracusa. Lucid: A practical, lightweight deep learning solution for ddos attack detection. *IEEE Transactions on Network and Service Management*, 17(2):876–889, 2020.

[9] Yebo Feng and Jun Li. Toward explainable and adaptable detection and classification of distributed denial-of-service attacks. In *International Workshop on Deployable Machine Learning for Security Defense*, pages 105–121. Springer, 2020.

[10] Romain Fontugne, Pierre Borgnat, Patrice Abry, and Kensuke Fukuda. MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking. In *Proceedings of the ACM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, 2010.

[11] Irom Lalit Meitei, Khundrakpam Johnson Singh, and Tanmay De. Detection of ddos dns amplification attack using classification algorithm. In *Proceedings of the International Conference on Informatics and Analytics*, pages 1–6, 2016.

[12] Chris Misa, Ramakrishnan Durairajan, Arpit Gupta, Reza Rejaie, and Walter Willinger. Leveraging prefix structure to detect volumetric ddos attack signatures with programmable switches. In *Security and Privacy*. IEEE, 2024 (to appear).

[13] José Jair Santanna, Roland van Rijswijk-Deij, Rick Hofstede, Anna Sperotto, Mark Wierbosch, Lisandro Zambenedetti Granville, and Aiko Pras. Booters—an analysis of ddos-as-a-service attacks. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 243–251. IEEE, 2015.

[14] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1:108–116, 2018.

[15] Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, and Ali A Ghorbani. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *computers & security*, 31(3):357–374, 2012.

[16] Tong Anh Tuan, Hoang Viet Long, Le Hoang Son, Raghvendra Kumar, Ishaani Priyadarshini, and Nguyen Thi Kim Son. Performance evaluation of botnet ddos attack detection using machine learning. *Evolutionary Intelligence*, 13(2):283–294, 2020.

[17] Xiaoyong Yuan, Chuanhuang Li, and Xiaolin Li. Deepdefense: identifying ddos attack via deep learning. In *2017 IEEE international conference on smart computing (SMARTCOMP)*, pages 1–8. IEEE, 2017.