Enhancing DDoS Detection: A Novel Real-World Dataset for Generalizable Cross-Domain Training

Balázs Nagy, Péter Orosz, Katica Bozsó, Tamás Skopkó
Dept. of Telecommunications and Artificial Intelligence
Budapest University of Technology and Economics
Budapest, Hungary
bnagy@tmit.bme.hu, orosz@tmit.bme.hu, katica.bozso@gmail.com, skopko@tmit.bme.hu

***Abstract*—Researchers working on DDoS detection encounter significant challenges, primarily due to the lack of robust datasets. Existing publicly available datasets are mostly synthetic, mixed, or severely limited. For example, the most cited studies since 2020 have relied on an aggregation of three synthetic datasets, none of which includes up-to-date YouTube and streaming traffic that constitute a significant portion of end-user data. This paper introduces four diverse datasets under the name SCLDDOS2024, collected from 8 Aug, 2022 to 26 Apr, 2024 on various commercial networks using uniform methods and tools. These datasets can significantly enhance DDoS research by supporting and validating the generalization of detection models. The datasets were collected using commercially available DDoS protection solutions and, although not optimal for all research purposes due to real-world data collection constraints, our analysis shows that even simple machine learning models, when paired with effective feature engineering, can achieve high accuracy with SCLDDOS2024. This work highlights the potential of SCLDDOS2024 to transform DDoS detection research by providing richer, more diverse real-world data.**

***Index Terms*—DDoS detection, Intrusion detection and prevention, DDoS, DDoS dataset, Machine learning, Security automation**

I. Introduction

Researchers working on DDoS detection face multiple challenges from which we believe the lack of corpus is the most dire. The publicly available datasets applied to validate current research are primarily synthetic, mixed, or extremely limited. To illustrate this problem, the most cited research since 2020 [1] built on the aggregation of three synthetic datasets, from which neither contained traces of real modern datacenter traffic. In this paper, we introduce four different datasets collected over 21 months from Q3'23 simultaneously in vast, different, commercial datacenter networks with the same methods and tools. These datasets can enhance DDoS research in multiple ways, supporting and validating the models' generalization: How will a model perform if deployed in a network different from the one it was developed? In addition, to publish our datasets, we highlight that they were collected by a commercially available DDoS protection solution. Therefore, we must acknowledge that our datasets are not optimal for all research-related purposes due to real-world constraints, such as the inability to capture all traffic in a network with peak incoming download throughput of 800Gbps for extended periods, for both technical and privacy reasons.

In this paper, we introduce SCLDDOS2024 incorporating four novel and diverse datasets for network traffic analysis.

II. Related Works

To understand the dataset problem the researchers face, we overview the most cited research of the past 4 years: LUCID: A Practical, Lightweight Deep Learning Solution for DDoS Attack Detection [1]; this research has multiple significant merits: it has a very innovative, efficient, and practical approach to DDoS detection. LUCID offers a valuable solution for enhancing cybersecurity in various network environments by addressing resource constraints and the need for unsupervised learning. The authors confirmed early on in their research that using different datasets for training and validation was going to produce poor results:

> *In an initial design, the model was trained and validated on the ISCX2012 dataset, producing high-accuracy results. However, testing the model on the CIC2017 dataset confirmed the generally held observation that a model trained on one dataset will not necessarily perform well on a completely new dataset.*

To battle this problem, the authors combined the three highest quality, publicly available datasets (ISCX2012 [2], CIC2017 [3], CSECIC2018 [4]) into one extensive dataset separating 10% of each for testing. The authors argue that a single dataset does not contain enough data for model training.

CSE-CIC [4], [5], [3], [2], [6] datasets are widely used for evaluating intrusion detection systems (IDS) and offer a comprehensive collection of network traffic data and associated labels for machine learning tasks. Each network traffic sample is meticulously labeled as either a type of attack or as benign traffic, facilitating supervised learning tasks. These are extremely valuable datasets for the scientific community and are probably the best general-purpose datasets available. However, there are a few drawbacks. These datasets are entirely synthetic, have a low data rate compared to real-world attacks, and the benign traffic is modeled after human user patterns; thus, they lack data-center server, cloud and IoT behaviors. Additionally, complex modern application-specific patterns are entirely missing, and the incorporated attacks usually do not achieve real denial-of-service on the victims.

978-3-903176-75-1 ©2025 IFIP

The authors of [1] combined three of these datasets to improve the overall data diversity.

DARPA 2000 Dataset [7] is one of the oldest and most well-known DDoS datasets. Nevertheless, its benign and attack patterns are highly outdated by today. The synthetic dataset consists of unencrypted, obsolete protocols, low-rate, obsolete attacks, no modern complex application traffic, no end-to-end encryption, etc. Alternative datasets offer more up-to-date traffic patterns.

CAIDA DDoS Attack 2007 [8] is a real-world collected dataset but spans over one hour only and contains no regular traffic. Its attack patterns are outdated and limited in both size and complexity.

MAWI-WIDE [9] is the most robust dataset available for research. This dataset represents the continuous capture of the transit traffic of the MAWI group from *2000* to *today*, constituting an enormous dataset with extreme historical significance and research potential. The dataset contains only a very narrow selection of traffic patterns limited to the applications present on this network, and only occasional attacks are present (the dataset is unbalanced).

BOĞAZIÇI University DDoS Dataset [10] is an intriguing dataset containing the *real* regular traffic of 4000 users on an enterprise-grade network. The benign traffic contains modern traffic patterns but is limited to end-user-type traffic. The attacks are especially limited in complexity and variation. Due to its age and the inclusion of real traffic, it is a valuable dataset for DDoS research; for optimal results, it should be combined with other datasets.

The DoS/DDoS Attack Dataset of 5G Network Slicing [11] is a distinctive dataset focusing on mobile network traffic including both DDoS and regular traffic. Although synthetic, it is unparalleled due to the unique characteristics of mobile network traffic. The distinction between access and mobile traffic is a double-edged sword; while it renders this dataset indispensable in 5G DDoS research, it constrains its application in other access-related research areas.

There are a few other datasets that we do not discuss in this paper due to the page length constraint. Nevertheless, we put a lot of consideration into showcasing the most important datasets.

III. The Proposed Dataset

The proposed datasets (see Table I) are unique in multiple aspects. This section will describe these datasets in detail and showcase their benefits. SCLDDOS2024 is publicly available at [12].

TABLE I: Comparison of the SCLDDOS2024 datasets

	Set A	**Set B**	**Set C**	**Set D**
Number of users	30k	600k	60k	200k
Transit	40GbE	800GbE	200GbE	80GbE
Non-transit UL	80GbE	1600GbE	400GbE	0
Server type users	30k	10k	60k	0

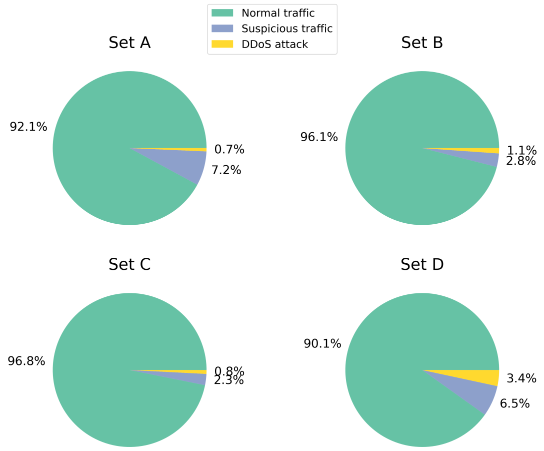

Fig. 1: Event type label distribution for all datasets

The most commonly used datasets are synthesized by researchers. In contrast, our dataset is collected in core networks of real-world data center providers. Many datasets are generated in controlled environments, which may not capture the full complexity of real-world network traffic and attacks.

The SCLDDOS2024 datasets are collected in 4 networks with vast differences in purpose and topology. Dataset A is collected in the network of a semi-public DCN hosting multiple large streaming services, and the B set is collected in the core network of a large regional internet service provider with mostly small users using mainly client services, set C collected in a public DCN and set D is collected in the core of a small regional provider.

The datasets were collected for the most extended timespan; while most public datasets were collected for days, our datasets were collected between 6-30 months per dataset.

They contain the richest attack patterns, containing more patterns than the discussed datasets combined. The vast transit networks between the endpoints transform the attack and regular traffic patterns. DDoS flow characteristics mutate during an attack. Large DDoS-es significantly differ from the small ones: flow characteristics change non-linearly. The frequency and length of attacks vastly differ from synthetic datasets. Some of the attacks in the dataset are extremely high volume, exceeding 300Gbps, showcasing the effect of network deterioration on concurrent regular traffic.

SCLDDOS2024 incorporates much borderline/anomalous traffic that is not necessarily an attack but certainly not normal user activity, see Fig. 1. Synthetic datasets almost entirely lack this kind of traffic.

A. Network measurement uncertainty under load

Network measurement uncertainty under load can be tricky because the commonly used built-in data collection functions usually offer best-effort services. For example, Cisco NetFlow, a widely used data gathering tool, is known to operators to become highly unreliable over 70% of uplink load. This observation holds true, albeit to a lesser extent, for 1:n network mirroring as well. The only lossless solution is passive optical tapping with a prohibitively high total cost of ownership

(TCO); hence, it is rarely used. Inline data gathering with lossless capture hardware is also available, but is generally discouraged by network operators due to its impact on network reliability; integrating additional equipment increases the risk of a catastrophic network blackout in case of failure.

Therefore, we propose that instead of seeking better data-gathering tools, we should accept and incorporate noisy data into our datasets as an essential requirement for evaluating algorithms and methods.

B. Data pre-processing pipeline

The data was gathered, which is the output of a commercial DDoS detection system. The methods, algorithms employed by this system are described in [13], [14]. The data collected using a uniform method, only minor patches had been applied to the data collection pipeline. During the period of data set creation, the system was periodically, manually evaluated for the correctness of the labels. Our capture card(s) are connected to the network's core routers and receive the network's mirrored ingress traffic on a 1:n port mirror. The packets are truncated on the router for maximal performance. A single 2x100GbE FPGA card can handle roughly 18x100GbE ingress traffic. All ingress traffic of the provider is mirrored: IP transit, peering, and IPx. The raw network packets are then collected into flows by an FPGA NIC with custom firmware [13], [14]. This FPGA card aggregates and collects the flows.

1) Traffic aggregation method: The most commonly used datasets employ flow-level aggregation, which, based on our decade-long experience, is not the ideal aggregation scheme for DDoS detection in real world scenarios. Instead, we opted for endpoint-level aggregation, which we argue is more effective. While theoretically flow-level aggregation can be losslessly reduced to endpoint-based aggregation - arguably making it a superior method - in practice, the most widely used aggregation tools compromise precision in various areas to facilitate flow-level aggregation. For example, most Cisco routers have a 5-minute aggregation period on NetFlow (2 minutes for high-end equipment). The continuous relevance of IP spoofing also makes flow-level aggregated data extremely messy; a small-sized DDoS can generate 300 *million* flow records in a minute, which, in practice, also should cause data loss somewhere in the processing pipeline. The information content of a spoofed flow with a few packets over 5 minutes is also highly minimal in DDoS detection.

We propose that processing power saved by endpoint-based aggregation should be applied to increase precision in other areas, such as sampling frequency. High-frequency sampling has many benefits, of which the greatest and almost mandatory is the reduced detection lag. Suppose the traffic is sampled with a 5-minute period. In that case, an average 2.5-minute detection lag is induced, which is unacceptable for most network operators, especially operators hosting critical services or operators with strict service level agreements. Accordingly, our datasets employ endpoint-level aggregation with 1s sampling. Also, endpoint-level aggregation makes a straightforward and very effective filtering option, discarding sub-threshold input traffic endpoints. Most services can handle a 50Mbps DDoS with ease. This could not be done for flow level aggregation because we have no idea how much harm a 5-packet flow causes. This rule also reduces the overall noise significantly because most normal network tasks require minimal throughput.

Our data has a feature called attack vectors, which is generated by a simple hardware-based DDoS detection algorithm. This algorithm measures the incoming and outgoing protocol traffic for protocols commonly used for DDoS, such as DNS and NTP. These attack vectors can be detected by elementary statistics, e.g., the incoming DNS responses are more numerous than the outgoing requests.

C. Upsides of these datasets

- Rich real-world patterns, extreme variance between attacks (10Mbps to 300Gbps)
- Highest number of users of all datasets
- Edge case, suspicious traffic (traffic that could only be categorized as malicious or normal by evaluating it with the context)
- Four datasets from different networks with the exact same collection method
- Port mirror saturates when high throughput DDoS attack is in progress, noisy data
- Highest number of DDoS attacks from any dataset

D. Limitation and downsides of our datasets

There are a few downsides to these datasets. These drawbacks mostly originate from the solution's commercial nature (cost saving). In this section, we list these drawbacks:

- No raw traces, simply the volume of data is too much for continuous capture
- Two features were generated by proprietary algorithms, which are removed from the datasets
- The data is aggregated per endpoint
- Low rate endpoint data is discarded at the FPGA level
- Port mirror saturates when high throughput DDoS attack is in progress, noisy data

E. Description of data

The datasets contain several fields. Some of them are directly used as features of the model, others can to be processed to create new features. Table II summarizes 'events' dataset fields.

The 'components' dataset has very similar fields but every record contains the data for a related component. The *Attack ID* field is used to correlate the events and components sets. The *Detect count* field of components set is the sequence number of the component itself, only unique per *Attack ID*.

Anonymization of *Victim IP* is applied globally throughout all the dataset partitions: unique IP addresses are taken sequentially, mapped to a positive integer number *X* and replaced to a label of *IP_X* format.

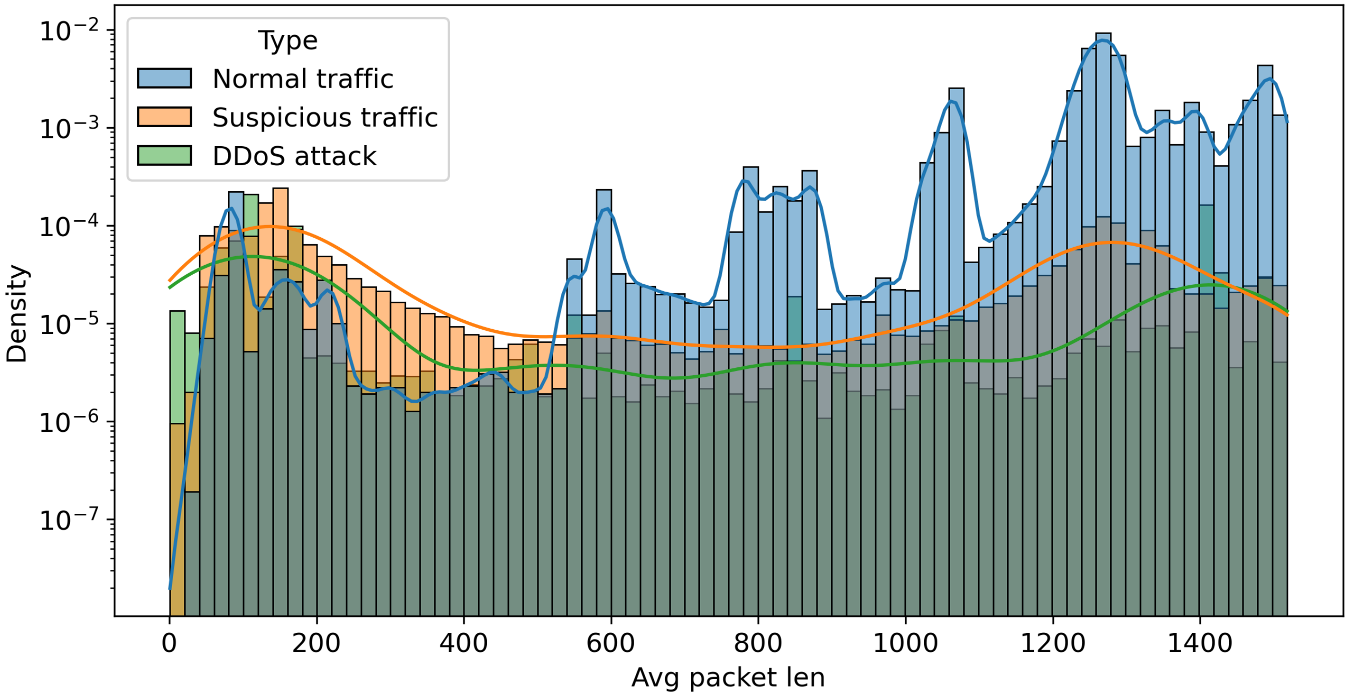

Fig. 2: Event type label distribution for all datasets

TABLE II: Description of the SCLDDOS2024 dataset features

Feature name	**Short description**	**Type**
Attack ID	Unique identifier of the event	categorical
Card	Network card name	categorical
Victim IP	Anonymized target IP	categorical
Port number	Destination port number	continuous, categorical
Attack code	Identifier of the most significant component	categorical
Detect count	Number of event components	continuous
Packet speed	Packet rate (pps)	continuous
Data speed	Data rate (bps)	continuous
Avg packet len	Average packet length (bytes)	continuous
Avg source IP count	Average source IP count of the components	continuous
Start time	Date and time of the event start	continuous
End time	Date and time of the event end	continuous
Type	Event category as recognized by the detector	categorical

F. Dataset properties and quality

Record count of set A, B and C is very close to each other: Set A contains 134770 events and 586642 components. Both set B and C contain 130000 events, but 1233449 components for B and 1247266 for C. Set D is a larger dataset with 437657 events and 2452610 components.

The dataset has no missing (null) values. The hardware accelerated data collection method ensures that no noise is filtrated into the data: all records have been filled properly, timestamps are generated on hardware basis as well. Data path of the collection is clear and thus there is no redundancy of events or components.

Number of entities in the three event classes (see *Type* attribute) are imbalanced: events classified as *Normal traffic* dominate all the sets, records in the two other categories are less about one order of magnitude . This is normal since there is no alert state in most of the time.

Most features show weak correlation to each other. *Data speed*, *Packet speed* and *Source IP count/Avg source IP* count are bound stronger. As well as *Duration* vs *Detect count*. Large volume traffic generally involves high packet speed and data rates. In this case it is more likely to identify more components for a longer event.

Distribution of *Data speed* for all the three categories show long tail form with some overlap, as well as for *Source IP count*.

Average packet lengths (see *Avg packet len* attribute) for the DDoS traffic are evenly distributed in the valid packet size range, as Fig. 2 depicts; current DDoS attacks typically consist of multiple attack vectors (i.e., attack types) with various packet sizes. In contrast, high-volume normal traffic is transmitted in large packets using the TCP transmission protocol, while its acknowledgement packets are typically small-sized, at about 100 bytes.

Since only start and end timestamps are recorded for the events and only start time for components, traffic shape cannot be reconstructed.

IV. Conclusion

In this paper, we have introduced SCLDDOS2024, a novel and comprehensive collection of four datasets aimed at addressing the critical need for robust, real-world data in DDoS detection research. Our datasets, collected over several years from different commercial networks using consistent methods, provide a rich variety of attack patterns and normal traffic, capturing the complexity and variance found in real-world scenarios. This diversity enhances the generalizability and robustness of machine learning models developed for DDoS detection.

Additionally, our research highlights the importance of real-world data in validating and improving DDoS detection models, moving beyond the limitations of synthetic datasets that have dominated the field.

Despite some constraints, such as the inability to capture all traffic continuously due to technical and privacy considerations, SCLDDOS2024 represents a significant step forward. It offers a valuable resource for researchers aiming to develop and test more accurate and generalizable DDoS detection methods. By making SCLDDOS2024 publicly available, we hope to stimulate further advancements in the field of DDoS detection, ultimately leading to more secure and resilient network infrastructures.

Acknowledgement

We would like to thank our long term college Pál Varga from the Budapest University of Technology and Economics for their research and work on which this research was based on. This research is also based on the data and development of AITIA International Llc.

References

[1] R. Doriguzzi-Corin, S. Millar, S. Scott-Hayward, J. Martínez-del Rincón, and D. Siracusa, “Lucid: A practical, lightweight deep learning solution for ddos attack detection,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 876–889, 2020.

[2] Institute for Cybersecurity. ISCXIDS2012. [Online]. Available: https://www.unb.ca/cic/datasets/ids.html

[3] ——. Intrusion detection evaluation dataset (CIC-IDS2017). [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html

[4] ——. CSE-CIC-IDS2018. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html

[5] ——. DDoS evaluation dataset (CIC-DDoS2019). [Online]. Available: https://www.unb.ca/cic/datasets/ddos-2019.html

[6] ——. ISCX NSL-KDD dataset 2009. [Online]. Available: https://www.unb.ca/cic/datasets/nsl.html

[7] DARPA. 1998 DARPA Intrusion Detection Evaluation Dataset. [Online]. Available: https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset

[8] Center for Applied Internet Data Analysis. CAIDA, Data Sharing. [Online]. Available: http://www.caida.org

[9] Working Group. MAWI Working Group Traffic Archive. [Online]. Available: https://mawi.wide.ad.jp/mawi/

[10] D. Erhan, “Boğaziçi university ddos dataset,” 2019. [Online]. Available: https://dx.doi.org/10.21227/45m9-9p82

[11] M. S. Khan, B. Farzaneh, N. Shahriar, and M. M. Hasan, “Dos/ddos attack dataset of 5g network slicing,” 2023. [Online]. Available: https://dx.doi.org/10.21227/32k1-dr12

[12] BME SmartCom Lab and AITIA. SCLDDoS2024. [Online]. Available: https://shorturl.at/ti9ev

[13] B. Nagy, P. Orosz, T. Tóthfalusi, L. Kovács, and P. Varga, “Detecting ddos attacks within milliseconds by using fpga-based hardware acceleration,” in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 2018, pp. 1–4.

[14] B. Nagy, P. Orosz, and P. Varga, “Low-reaction time fpga-based ddos detector,” in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 2018, pp. 1–2.