



DDoS Attack Datasets: A Comprehensive Analytical Review of Representative Datasets

Dr. R. Anurekha

Assistant Professor

Department of Information Technology

Government College of Engineering, Erode – 638316, Tamil Nadu, India

Abstract: Distributed Denial-of-Service (DDoS) attacks remain a significant threat to the availability and reliability of Internet-based services. The increasing diversity of attack mechanisms, communication protocols, network architectures, and deployment environments has resulted in the development of numerous datasets for DDoS and network-security research. However, these datasets differ substantially in their provenance, traffic-generation methodology, network environment, attack coverage, traffic representation, labeling, temporal characteristics, and accessibility. Consequently, datasets containing DDoS traffic cannot be considered equivalent, and selecting an appropriate dataset requires an understanding of these differences. This paper presents a comprehensive analytical review of representative publicly available datasets relevant to DDoS research. Rather than attempting to catalogue every dataset reported in the literature, the study applies a structured screening methodology to identify representative non-IoT datasets and distinguish dedicated DDoS datasets from broader network-security datasets containing DDoS traffic. Ten representative datasets are initially examined according to their origin, network environment, DDoS relevance, attack coverage, traffic representation, temporal characteristics, and principal limitations. Three complementary datasets—CAIDA DDoS Attack 2007, CICDDoS2019, and the Boğaziçi University DDoS Dataset—are subsequently selected for detailed analysis because they represent distinct data-collection conditions: an observed real Internet DDoS event, a dedicated controlled multi-vector DDoS environment, and a university campus network containing experimentally generated DDoS traffic alongside real background traffic. The comparative analysis demonstrates that the datasets examined in this study provide complementary forms of evidence rather than interchangeable representations of the DDoS threat. Observational datasets provide operational context but may be limited in event and attack diversity, whereas controlled datasets provide structured attack scenarios and documented experimental conditions but may not reproduce all characteristics of operational Internet traffic. The study further identifies variations in attack taxonomy, temporal representation, background traffic, ground truth, network context, and traffic representation. Based on these observations, requirements for future DDoS datasets are identified, including documented provenance, diverse attack scenarios, temporal annotations, representative benign traffic, reliable ground truth, packet and flow representations, and transparent documentation.

Keywords: Distributed Denial-of-Service, DDoS datasets, network traffic, cybersecurity, dataset analysis, network security, intrusion detection

1. INTRODUCTION

Distributed Denial-of-Service (DDoS) attacks attempt to impair the availability of network services by generating traffic or requests at a scale or rate that overwhelms a target system, network resource, or application. The distributed nature of many attacks makes the problem particularly challenging because traffic originating from multiple sources can collectively impose substantial load on a target while individual flows may resemble

legitimate communication. As dependence on Internet-based services continues to increase, DDoS attacks remain an important subject of network-security research.

The development and evaluation of DDoS analysis and detection techniques depend substantially on the availability of representative network-traffic datasets. Datasets provide the empirical basis for examining traffic characteristics, identifying attack patterns, studying temporal behavior, and evaluating analytical methods. However, the availability of a large number of datasets has introduced an additional methodological problem: researchers must determine whether a particular dataset actually represents the DDoS characteristics relevant to a given research question.

A dataset may contain a label named “DDoS” while having originally been developed as a general intrusion-detection dataset. Conversely, a dedicated DDoS dataset may provide detailed attack categories but be generated primarily within a controlled laboratory environment. Observational datasets may provide operational Internet traffic but contain only a limited number of documented events. Therefore, dataset size, number of attack labels, or the presence of a DDoS class alone is insufficient to establish analytical suitability.

Several widely used datasets illustrate these differences. CICIDS2017 is a general intrusion-detection dataset containing benign traffic and several attack categories, including DDoS. CSE-CIC-IDS2018 similarly contains seven broad attack scenarios, including DDoS, but was designed as a broader cyber-defense dataset rather than exclusively as a DDoS collection. In contrast, CICDDoS2019 was specifically developed as a DDoS evaluation dataset and includes multiple documented DDoS attack scenarios. CAIDA DDoS Attack 2007 represents another fundamentally different category because it contains anonymized traffic traces associated with an actual DDoS event observed on the Internet.

These differences motivate a systematic dataset-oriented analysis. The purpose of this paper is therefore not to compare DDoS detection algorithms or machine-learning models, but to examine the datasets themselves and identify the characteristics that determine their analytical value. The study deliberately excludes IoT- and IIoT-specific datasets from the principal comparison so that the analysis remains focused on conventional Internet and network-based DDoS traffic.

The principal contributions of this study are as follows. First, a structured screening framework is proposed for distinguishing dedicated DDoS datasets from broader network-security datasets containing DDoS traffic. Second, representative non-IoT datasets are comparatively examined using common analytical dimensions. Third, three complementary datasets are examined in detail to illustrate differences between observed Internet attacks, controlled multi-vector DDoS scenarios, and experimentally generated attacks occurring alongside real background traffic. Finally, the study identifies requirements for future DDoS datasets intended to support reproducible and cross-dataset research.

II. MOTIVATION

INCREASING DIVERSITY OF DDoS DATASETS

The DDoS dataset landscape has evolved considerably over time. Earlier datasets were frequently associated with individual attack events or relatively small controlled experiments, whereas later datasets have attempted to incorporate multiple attack vectors, realistic background traffic, richer traffic representations, and more explicit labeling.

CAIDA DDoS Attack 2007, for example, provides approximately one hour of anonymized traffic traces associated with a specific DDoS event observed on August 4, 2007. CICDDoS2019, in contrast, was designed specifically as a DDoS evaluation dataset and contains multiple contemporary DDoS attack scenarios. These datasets therefore provide fundamentally different forms of evidence even though both are used in DDoS research. The evolution of datasets has increased data availability but has also made direct comparison more difficult. A large flow-based dataset, a short packet trace from a real incident, and a controlled multi-vector dataset may all be described as “DDoS datasets,” while their analytical properties are substantially different.

AMBIGUITY IN DDoS DATASET CLASSIFICATION

A major challenge in reviewing DDoS datasets is the absence of a consistently applied distinction between dedicated DDoS datasets and general intrusion datasets containing DDoS traffic.

A general intrusion dataset may contain DDoS as one class alongside brute-force attacks, web attacks, scanning, botnet activity, infiltration, or other threats. Such datasets remain useful for broader network-security studies, but they should not automatically be treated as dedicated DDoS datasets.

The distinction is important because dataset construction affects the conclusions that can reasonably be drawn from the data. A dataset designed specifically around DDoS may provide detailed attack categories and attack-generation information, whereas a general intrusion dataset is intended to represent a broader security environment.

Accordingly, this study proposes two descriptive categories:

Primary DDoS dataset: a dataset in which DDoS constitutes a principal objective or a clearly defined central component of dataset construction.

Secondary DDoS-relevant dataset: a broader network-security or traffic dataset in which DDoS is represented as one of several attack categories or as a documented subset.

The classification is based on the stated purpose and documented construction of the dataset rather than on the frequency with which the dataset is used in DDoS research.

VARIATION IN NETWORK ENVIRONMENTS

DDoS traffic can be observed across different network environments, including Internet backbones, ISP networks, enterprise networks, campus networks, data centers, and software-defined networks. Consequently, traffic characteristics may be influenced by the underlying environment as well as by the attack mechanism.

CAIDA DDoS Attack 2007 represents an Internet measurement environment, whereas the Boğaziçi University dataset represents traffic captured from a university campus network. CICDDoS2019 represents a controlled experimental environment designed for DDoS evaluation. SDN-oriented datasets such as InSDN and HLD-DDoS introduce another dimension because their traffic is associated with software-defined networking environments.

These differences should be explicitly considered before datasets are compared or combined.

DIVERSITY OF DDoS ATTACK CHARACTERISTICS

DDoS attacks are not represented by a single traffic pattern. High-rate flooding, protocol-oriented attacks, reflection/amplification attacks, SYN-based attacks, slow-rate attacks, and application-layer attacks can produce substantially different traffic characteristics.

CICDDoS2019 illustrates this diversity by including NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag, WebDDoS, SYN, and TFTP attacks in its training scenarios. The testing scenarios include PortScan together with NetBIOS, LDAP, MSSQL, UDP, UDP-Lag, and SYN attacks.

In contrast, CAIDA DDoS Attack 2007 represents a specific historical event, while HLD-DDoS focuses on TCP, UDP, and ICMP DDoS attacks with high- and low-rate variations in an SDN environment.

Consequently, the number and type of attack categories should be considered explicitly in dataset analysis.

VARIATION IN DATA REPRESENTATION

Datasets differ not only in the attacks they contain but also in how traffic is represented. Some provide packet captures, some provide flow-level records, and some provide both.

Packet traces preserve detailed temporal and protocol information, while flow-level records provide aggregated descriptions of communication sessions. CICDDoS2019 provides PCAP traffic together with derived labelled flow records. UGR'16 provides real anonymized NetFlow traces collected at a Tier-3 ISP over four months.

The representation determines the types of analysis that can be performed. Packet traces can support detailed protocol and temporal analysis, whereas flow records facilitate large-scale traffic characterization. Therefore, a dataset review should examine the underlying traffic representation rather than relying only on the number of records or features.

NEED FOR A STRUCTURED DATASET ASSESSMENT

The differences described above demonstrate the need for a structured approach to dataset assessment. A meaningful comparison should consider provenance, network environment, attack generation, attack diversity, benign traffic, temporal information, labeling, data representation, accessibility, and limitations.

The present study addresses this requirement through a multi-stage screening and comparative analysis framework.

III. METHODOLOGY

This study adopts a dataset-oriented methodology for identifying and analyzing representative DDoS datasets. The analysis is deliberately restricted to a manageable set of non-IoT datasets rather than attempting to enumerate every cybersecurity dataset that has been used in DDoS-related research.

The methodology consists of three principal stages. The first stage identifies and screens candidate datasets according to their documented DDoS relevance. The second stage performs an initial comparative assessment of the retained datasets. The third stage selects three complementary datasets for detailed analysis.

DATASET IDENTIFICATION

Candidate datasets were identified from official dataset repositories, dataset publications, institutional sources, and established cybersecurity research resources. Preference was given to datasets with identifiable provenance and publicly documented collection or generation methodologies.

The search was centered on combinations of terms such as “DDoS dataset,” “distributed denial-of-service dataset,” “DDoS network traffic dataset,” “DDoS attack trace,” and “DDoS intrusion dataset.” Dataset documentation and original publications were examined to establish the purpose, environment, traffic representation, attack coverage, and availability of each candidate.

The initial selection focused on conventional Internet and network-based DDoS environments. Datasets whose principal purpose was IoT or IIoT security were excluded from the main comparison to avoid introducing domain-specific traffic characteristics that could confound the general DDoS analysis.

Derived or repackaged datasets were also not counted as independent datasets when their underlying traffic originated from an already included benchmark.

DDoS SUITABILITY SCREENING

Each candidate dataset was assessed according to its relationship with DDoS attacks.

A Primary DDoS dataset is one in which DDoS constitutes a principal objective or a clearly defined central component of dataset construction. A Secondary DDoS-relevant dataset is a broader network-security or traffic dataset in which DDoS is represented as one of several attack categories or as a documented subset.

This distinction prevents general intrusion datasets from being treated as equivalent to dedicated DDoS datasets.

COMPARATIVE ASSESSMENT DIMENSIONS

The initial analysis considers the following dimensions: dataset origin and year, network environment, DDoS classification, attack coverage, traffic representation, benign-traffic context, temporal coverage, labeling and ground truth, accessibility and principal limitations. The objective is descriptive and analytical rather than algorithmic. No attempt is made to rank datasets according to detection accuracy, machine-learning performance, or an overall quality score.

SELECTION OF REPRESENTATIVE DATASETS FOR DETAILED ANALYSIS

Three datasets were selected for detailed examination according to their ability to represent substantially different DDoS data-generation conditions.

CAIDA DDoS Attack 2007 represents an observed real Internet event. CICDDoS2019 represents a dedicated controlled multi-vector DDoS environment. The Boğaziçi University DDoS Dataset represents a university campus network in which resource-depletion DDoS attacks were generated while background traffic from real Internet users was observed.

The selection is therefore complementary rather than a ranking of dataset quality.

IV. INITIAL COMPARATIVE ANALYSIS OF REPRESENTATIVE DATASETS RELEVANT TO DDoS RESEARCH

INITIAL DATASET SET

Ten representative non-IoT datasets were retained for the initial screening: CAIDA DDoS Attack 2007, DDoS 2016, CICIDS2017, CSE-CIC-IDS2018, CICDDoS2019, Boğaziçi University DDoS Dataset, InSDN, UGR'16, SDN-SlowRate-DDoS and HLD-DDoSNDN

The selected datasets cover observed Internet traffic, controlled DDoS experiments, university operational traffic, general intrusion-detection collections, ISP flow data, and software-defined networking environments.

Table 1. Initial screening of representative datasets relevant to DDoS research

Dataset	Year	Environment	Classification	Principal DDoS relevance	Main limitation
CAIDA DDoS Attack 2007	2007	Internet measurement environment	P	Observed real DDoS event	Single historical event and limited attack diversity
DDoS 2016	2016	Controlled network	P	UDP flood, Smurf, HTTP flood, SIDDOS	Controlled environment
CICIDS2017	2017	General network environment	S	DDoS within broader attack taxonomy	Not DDoS-specific
CSE-CIC-IDS2018	2018	Network/cloud environment	S	Explicit DDoS scenario	Broad IDS scope
CICDDoS2019	2019	Dedicated controlled DDoS environment	P	Multiple documented DDoS vectors	Controlled traffic generation
Boğaziçi University DDoS	2020	University campus network	P	Resource-depletion DDoS with real background traffic	Limited attack/environment scope
InSDN	2020	SDN	S	DDoS within broader SDN attack scenarios	Architecture-specific
UGR'16	2016, 2018	Tier-3 ISP	S	Long-duration ISP NetFlow containing documented network attacks	General network IDS scope
SDN-SlowRate-DDoS	2023	Physical SDN testbed	P	Slow-rate DDoS attacks	SDN and slow-rate specific
HLD-DDoS	2024	SDN	P	High- and low-rate TCP, UDP, and ICMP DDoS	SDN-specific

P: Primary DDoS dataset.

S: Secondary DDoS-relevant dataset.

DDoS 2016 is associated with the study by Alkasassbeh et al., which describes UDP flood, Smurf, HTTP flood, and SIDDOS attack traffic. CICIDS2017 and CSE-CIC-IDS2018 contain DDoS as part of broader attack taxonomies.

CICDDoS2019 was explicitly developed as a DDoS evaluation dataset and provides PCAP traffic together with labelled flow records. The Boğaziçi dataset combines attack traffic with attack-free traffic from more than 4,000 real Internet users.

InSDN is a broader SDN intrusion dataset rather than a DDoS-only collection. UGR'16 provides real anonymized NetFlow traces collected at a Tier-3 ISP for four months and was developed for network IDS evaluation.

The SDN-SlowRate-DDoS dataset was introduced through a physical SDN security framework study and contains experiments involving slow-rate DDoS attacks on a physical testbed. HLD-DDoS focuses on TCP, UDP, and ICMP DDoS attacks with high- and low-rate variations and contains 71 statistically qualified traffic features.

ANALYTICAL OBSERVATIONS FROM INITIAL SCREENING

The initial screening demonstrates that DDoS-relevant datasets can broadly be differentiated according to how attack traffic and background traffic are obtained. Real-event datasets such as CAIDA provide observations of naturally occurring Internet attacks. Controlled datasets such as CICDDoS2019 provide greater control over attack categories and experimental conditions. The Boğaziçi University dataset combines experimentally generated attacks with background traffic from real users.

The distinction is analytically important. Real-world traffic provides contextual information from an operational network, but researchers have limited control over the attack conditions. Controlled experiments allow specific attack conditions to be generated and documented, but the resulting traffic may not reproduce all characteristics of operational Internet attacks.

Attack diversity is another important dimension. CICDDoS2019 includes numerous DDoS categories, whereas CAIDA represents a particular historical event. These datasets therefore support different research questions.

General intrusion datasets should also be treated differently from dedicated DDoS collections. CICIDS2017, CSE-CIC-IDS2018, and InSDN provide DDoS traffic within broader cybersecurity environments. Such datasets can be useful when DDoS is considered as one component of a larger network-security problem, but their broader scope should be acknowledged.

The temporal dimension is particularly important for DDoS analysis. Short event-oriented datasets provide information about individual attack periods, whereas long-duration datasets such as UGR'16 provide a broader view of network activity over extended periods. UGR'16 was constructed from real anonymized NetFlow data collected at an ISP over four months.

V. DETAILED ANALYSIS OF THE THREE SELECTED DATASETS

CAIDA DDoS ATTACK 2007

DATASET ORIGIN AND PURPOSE

The CAIDA DDoS Attack 2007 dataset is associated with an actual DDoS attack observed on August 4, 2007. CAIDA reports approximately one hour of anonymized traffic traces covering the period from 20:50:08 UTC to 21:56:16 UTC.

The dataset differs fundamentally from controlled DDoS testbeds because it records traffic associated with an observed Internet event rather than a scenario created solely for dataset generation.

NETWORK ENVIRONMENT

The dataset represents traffic observed in an Internet measurement environment. This provides operational context that differs from an isolated laboratory network.

At the same time, the network environment corresponds to a particular historical period. Network architectures, application usage, traffic volumes, and attack techniques have changed substantially since 2007. Therefore, historical representativeness should not be confused with contemporary representativeness.

TRAFFIC REPRESENTATION

The dataset provides anonymized traffic traces. The availability of packet traces enables temporal and protocol-level examination of the observed event, while anonymization limits direct interpretation of endpoint identities and some contextual information.

The dataset should consequently be treated as an anonymized observational trace rather than as a fully identified operational network dataset.

TEMPORAL CHARACTERISTICS

The temporal structure is a major analytical property because the dataset corresponds to a clearly documented observation interval. This allows traffic behavior to be studied over the recorded event period.

The relatively short observation period is also a limitation. A single event cannot represent the temporal diversity of DDoS attacks across different days, network conditions, campaigns, and environments.

GROUND TRUTH AND INTERPRETATION

The association with a documented DDoS event provides event-level context. However, event-level ground truth should not be interpreted as a complete taxonomy of all possible DDoS mechanisms.

The dataset is therefore most appropriately interpreted as evidence from a specific real DDoS event rather than as a comprehensive collection of DDoS attack vectors.

PRINCIPAL STRENGTHS AND LIMITATIONS

The principal analytical value of CAIDA DDoS Attack 2007 lies in its observed Internet origin and historical operational context. Its limitations include its historical age, concentration on a single event, anonymization, and limited event diversity.

The dataset consequently provides an important reference when the research question concerns the characteristics of traffic observed during an actual Internet DDoS incident.

CICDDoS2019

DATASET ORIGIN AND PURPOSE

CICDDoS2019 was developed by the Canadian Institute for Cybersecurity as a dedicated DDoS evaluation dataset. The dataset was designed to provide benign and contemporary DDoS traffic for evaluating DDoS detection systems.

Unlike CAIDA, CICDDoS2019 was constructed under controlled experimental conditions. This permits attack scenarios to be defined and traffic to be organized according to documented attack categories.

ATTACK COVERAGE

The training scenarios include NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag, WebDDoS, SYN, and TFTP. The testing scenarios include PortScan together with NetBIOS, LDAP, MSSQL, UDP, UDP-Lag, and SYN.

This broad attack coverage distinguishes CICDDoS2019 from event-specific datasets such as CAIDA. It allows different documented attack mechanisms to be examined within a common dataset-generation framework.

NETWORK AND EXPERIMENTAL ENVIRONMENT

The dataset was generated using a controlled experimental network involving multiple systems and documented attack scenarios. Such control provides consistency across attack categories and makes it possible to associate observed traffic with experimental conditions.

However, controlled traffic generation cannot reproduce every characteristic of naturally occurring Internet attacks. Background traffic, routing conditions, source diversity, user behavior, and network-scale effects may differ from those observed during operational attacks.

DATA REPRESENTATION

CICDDoS2019 provides captured network traffic in PCAP form together with derived labelled flow records. The official dataset documentation states that flow records were generated using CICFlowMeter-V3 and labelled using information including timestamps, source and destination IP addresses, ports, protocols, and attack categories.

This dual representation is analytically useful because researchers can examine packet-level information or derive alternative flow-level representations from the captured traffic.

TEMPORAL AND LABEL CHARACTERISTICS

Because attack scenarios are experimentally generated, the dataset provides a structured relationship between traffic and documented attack categories. This facilitates attack-type-level analysis.

However, the labels describe the experimental scenarios and should not automatically be interpreted as universal representations of how corresponding attacks behave across all operational networks.

PRINCIPAL STRENGTHS AND LIMITATIONS

CICDDoS2019 provides broad documented DDoS attack coverage, controlled experimental conditions, and packet/flow representations. Its principal limitation is the controlled nature of traffic generation.

The dataset therefore provides a useful common framework for examining differences among multiple documented DDoS mechanisms.

BOĞAZIÇI UNIVERSITY DDoS DATASET

DATASET ORIGIN AND PURPOSE

The Boğaziçi University Distributed Denial of Service Dataset was developed to provide DDoS traffic together with attack-free user traffic from an operational university network. The dataset includes attack traffic and attack-free traffic and was designed for network-based DDoS analysis.

The dataset therefore occupies an intermediate position between an entirely observational dataset and a fully isolated attack testbed: the attack traffic is experimentally generated, while the background traffic originates from real users in the campus network.

OPERATIONAL NETWORK CONTEXT

The dataset was collected at Boğaziçi University in Istanbul, Turkey. The dataset description states that attack-free traffic includes traffic from more than 4,000 real Internet users.

This characteristic is analytically significant because attack traffic is represented in the context of substantial legitimate network activity rather than as an isolated attack stream.

ATTACK GENERATION

The dataset contains resource-depletion DDoS attack traffic generated within the university network. The attack periods can be related to the experimentally generated attack traffic while legitimate background traffic continues to be observed.

This construction provides experimental control over the attack while retaining real background traffic.

TRAFFIC REPRESENTATION AND PRIVACY

The dataset is based on network traffic captured from the campus backbone environment. The original description indicates that payloads were removed and source IP addresses of legitimate users were anonymized to protect confidentiality.

The privacy-preserving treatment is important when interpreting endpoint information and illustrates a general trade-off between analytical detail and privacy protection in operational network datasets.

PRINCIPAL STRENGTHS AND LIMITATIONS

The primary analytical value of the Boğaziçi University dataset is its combination of experimentally generated DDoS traffic and background traffic from a university network containing more than 4,000 real Internet users.

Its principal limitation is the concentration of the data collection within one institutional environment and its focus on a particular set of resource-depletion attack scenarios. It should therefore not be interpreted as a representation of all DDoS mechanisms or all operational network environments.

VI. DATASET COMPARATIVE ANALYSIS

The three selected datasets represent different approaches to DDoS data generation. Their comparison is therefore more informative when based on analytical dimensions rather than on a single measure of dataset size.

Table 2: Detailed comparison of the three selected datasets

Dimension	CAIDA DDoS Attack 2007	CICDDoS2019	Boğaziçi University DDoS
Primary purpose	Observation of a real DDoS event	Dedicated DDoS evaluation	DDoS traffic with campus background traffic
Collection context	Observed Internet traffic	Controlled experimental traffic	Campus traffic with generated attacks and real background
Network environment	Internet measurement environment	Controlled experimental network	University campus network

Attack origin	Observed real event	Experimentally generated	Experimentally generated
Attack diversity	Event-specific	Multiple documented DDoS categories	Resource-depletion scenarios
Normal traffic context	Event-associated Internet traffic	Controlled benign traffic	Real user background traffic
Traffic representation	Anonymized packet traces	PCAP and labelled flow records	Captured campus network traffic
Temporal perspective	Short historical event	Multiple controlled scenarios	Attack periods within background traffic
Ground-truth context	Documented real event	Experimental scenario labels	Known attack-generation periods
Collection context	Operational observation	Controlled experiment	Operational background + controlled attack
Main analytical contribution	Real Internet event perspective	Multi-vector attack coverage	Real background traffic with generated DDoS
Principal limitation	Historical and event-specific	Controlled environment	Limited attack and environment scope

The comparison demonstrates that the three datasets are complementary. CAIDA provides an observational perspective, CICDDoS2019 provides experimental breadth, and the Boğaziçi University dataset combines generated attacks with real background activity.

COLLECTION CONTEXT AND EXPERIMENTAL CONTROL

One of the most important differences is the relationship between operational observation and experimental control. CAIDA provides evidence from an actual Internet event, whereas CICDDoS2019 provides a controlled environment in which attack categories are explicitly defined. The Boğaziçi University dataset combines experimentally generated DDoS traffic with background traffic from real users.

This distinction should be considered whenever datasets are compared. Controlled datasets can provide clearly defined attack scenarios, while operational datasets can provide contextual background traffic that is difficult to reproduce in a laboratory.

ATTACK DIVERSITY AND EVENT SPECIFICITY

CICDDoS2019 contains a broad collection of documented DDoS scenarios, whereas CAIDA focuses on a particular historical event.

The Boğaziçi University dataset occupies another position because it emphasizes resource-depletion DDoS attacks in combination with legitimate traffic from a university campus network.

Consequently, the datasets provide different forms of evidence and should not be treated as interchangeable representations.

PACKET-LEVEL AND FLOW-LEVEL ANALYSIS

The availability of packet-level traffic is important for detailed protocol and temporal analysis. CICDDoS2019 provides PCAP traffic and derived flow-level records, while CAIDA provides anonymized traffic traces.

Flow-level datasets such as UGR'16 provide another analytical perspective through long-duration NetFlow observations collected at an ISP.

Researchers should therefore determine whether their research question requires packet-level data, flow-level data, or both before selecting a dataset.

TEMPORAL CHARACTERISTICS

Temporal information is particularly important for DDoS analysis because an attack is a dynamic event rather than a static class. Attack onset, escalation, peak activity, persistence, and termination may provide information that is lost when traffic is reduced to independent records.

CAIDA provides a clearly bounded historical event, CICDDoS2019 provides multiple controlled attack scenarios, and the Boğaziçi University dataset permits examination of attack traffic in relation to simultaneous legitimate activity.

Temporal representation should therefore be considered an independent dataset-selection dimension.

VII. ANALYTICAL FINDINGS AND RESEARCH GAPS

The comparative study reveals several important characteristics of the DDoS dataset landscape.

First, none of the datasets examined in this study simultaneously provides extensive attack diversity, long-term operational traffic, multiple real-world attack events, detailed packet information, comprehensive ground truth, and broad network-environment coverage. The reviewed datasets generally emphasize some of these characteristics while providing less coverage of others.

Second, operational context remains difficult to combine with broad attack diversity. CAIDA provides evidence from an observed Internet event, while CICDDoS2019 provides multiple controlled attack scenarios. The Boğaziçi University dataset combines experimentally generated attacks with background traffic from real users.

Third, temporal information is not consistently represented across datasets. DDoS attacks are dynamic phenomena, yet many analytical representations reduce traffic to individual flow records. Such representations can be useful for statistical analysis, but they may obscure attack onset, burst structure, persistence, and changes in traffic intensity.

Fourth, attack taxonomy remains inconsistent. Different datasets use different names and groupings for attacks. CICDDoS2019 distinguishes multiple protocol- and service-oriented DDoS categories, while general IDS datasets use broader attack taxonomies.

Fifth, background traffic characteristics influence the interpretation of DDoS traffic. A dataset containing attack traffic without substantial legitimate activity does not provide the same context as a dataset in which attacks occur alongside legitimate users. The Boğaziçi University dataset illustrates this distinction through attack-free traffic involving more than 4,000 real Internet users.

Sixth, dataset age should be considered independently of dataset quality. CAIDA DDoS Attack 2007 remains historically valuable because it represents an observed Internet DDoS event, but its traffic reflects a network environment from 2007. More recent datasets may contain newer attack scenarios, but newer data are not automatically representative of all DDoS environments.

Seventh, accessibility is a practical characteristic that can change over time. CAIDA reported that the DDoS Attack 2007 dataset became publicly accessible in August 2025, although users with previous access could continue using the request-based channel. Consequently, dataset accessibility should be reported according to the status applicable at the time of research.

VIII. REQUIREMENTS FOR A COMPREHENSIVE DDOS DATASET

The comparative findings indicate that future DDoS datasets should be designed around several complementary requirements.

A comprehensive dataset should provide a clearly documented network environment and data-collection methodology. Researchers should be able to determine whether traffic originated from an operational network, a controlled testbed, a simulation, or a hybrid environment.

The dataset should also provide explicit information concerning attack generation. The attack type, protocol, source characteristics, target characteristics, attack duration, and attack-generation mechanism should be documented sufficiently to permit independent interpretation of the traffic.

Temporal information should be preserved wherever possible. Packet timestamps, attack start and termination periods, and relevant changes in traffic intensity are important for studying DDoS attacks as dynamic events.

The representation of legitimate traffic should receive equal attention. A DDoS dataset should not be evaluated solely according to the amount of attack traffic it contains. The characteristics and provenance of normal traffic determine the context in which attack behavior can be interpreted.

Where privacy permits, information concerning source and destination characteristics should be retained or appropriately anonymized. Anonymization is necessary in many operational datasets, but the resulting analytical

limitations should be clearly documented. The Boğaziçi University dataset provides an example in which legitimate-user source IP addresses were anonymized and packet payloads were removed.

A comprehensive dataset should ideally provide both packet-level and flow-level representations. Packet traces preserve detailed network information, while flow records provide an efficient representation for large-scale analysis. Providing both allows researchers to derive representations appropriate to different research questions. Finally, dataset documentation should clearly describe ground truth. Labels should be associated with documented attack-generation events wherever possible, and the relationship between attack periods and individual records should be explicitly specified.

IX. IMPLICATIONS FOR DDoS DATASET SELECTION

The findings of this study indicate that dataset selection should be driven by the research question rather than by dataset popularity or record volume.

When the objective is to examine DDoS behavior observed during a real Internet event, CAIDA provides an observational reference. When the objective is to examine differences among multiple documented DDoS mechanisms under controlled conditions, CICDDoS2019 provides a structured multi-vector dataset. When the objective is to examine generated DDoS attacks in the presence of substantial legitimate campus traffic, the Boğaziçi University dataset provides another form of evidence.

The three datasets should therefore be regarded as complementary sources rather than interchangeable alternatives. Combining evidence from datasets with different collection methodologies can provide a broader understanding of DDoS traffic, provided that differences in provenance, environment, labeling, and traffic representation are explicitly considered.

This observation also has implications for comparative experimental studies. Results obtained from one dataset should not automatically be generalized to all DDoS environments. A method or analytical observation derived from a controlled multi-vector dataset may behave differently when applied to observed Internet traffic, while observations from a single historical event may not capture the diversity represented in a modern multi-vector dataset.

X. FUTURE RESEARCH DIRECTIONS

Future DDoS dataset development should move toward collections that preserve both attack diversity and operational context. Long-duration datasets containing multiple attack events would be valuable because they could capture variations in attack intensity, background traffic, network conditions, and attack mechanisms.

Another important direction is preservation of temporal structure. Rather than treating traffic records as independent observations, future datasets could provide event-level annotations that explicitly identify attack onset, escalation, peak activity, persistence, and termination.

Cross-dataset standardization is also required. Consistent definitions of attack categories, labeling conventions, timestamp formats, traffic representations, and metadata would improve the ability to compare datasets without obscuring their underlying differences.

Future datasets should additionally document the relationship between attack traffic and legitimate traffic. The provenance of benign traffic, the number and diversity of legitimate users, and the network context in which background traffic was captured are important characteristics that are often less clearly documented than attack labels.

Emerging network architectures and application environments will require additional dedicated datasets. However, domain-specific datasets should be analyzed separately from general Internet DDoS datasets when architectural characteristics substantially influence traffic behavior. In particular, IoT and IIoT datasets can form a separate research category rather than being combined indiscriminately with conventional Internet DDoS datasets.

XI. CONCLUSION

This paper presented a comprehensive analytical review of representative datasets relevant to DDoS research. Rather than treating all datasets containing DDoS traffic as equivalent, the study established a distinction between dedicated DDoS datasets and broader network-security datasets containing DDoS scenarios.

Ten representative non-IoT datasets were initially screened according to their provenance, network environment, DDoS relevance, attack coverage, traffic representation, temporal characteristics, and limitations. The screening included observed Internet traffic, controlled DDoS experiments, university network traffic, ISP flow data, general intrusion datasets, and SDN-specific datasets.

Three datasets—CAIDA DDoS Attack 2007, CICDDoS2019, and the Boğaziçi University DDoS Dataset—were subsequently examined in detail because they represent complementary DDoS data-generation conditions. CAIDA provides an observational perspective based on a real Internet DDoS event. CICDDoS2019 provides a controlled environment containing multiple documented DDoS attack scenarios. The Boğaziçi University dataset combines experimentally generated DDoS traffic with legitimate background traffic from a university campus network containing more than 4,000 real Internet users.

The analysis demonstrates that DDoS dataset characteristics cannot be adequately described by dataset size or the number of attack labels alone. The origin of traffic, network environment, attack-generation mechanism, temporal coverage, background traffic, labeling methodology, privacy treatment, and traffic representation all influence the type of analysis that a dataset can support.

The study consequently concludes that the datasets examined provide complementary views of DDoS traffic rather than a universally interchangeable representation. A future comprehensive DDoS dataset should seek to combine diverse attack scenarios, representative background traffic, precise temporal information, reliable ground truth, packet and flow representations, documented collection methodology, and transparent metadata.

Such characteristics would improve reproducibility, facilitate cross-dataset analysis, and strengthen the validity of conclusions drawn from DDoS traffic research.

REFERENCES

- [1] CAIDA, “The CAIDA UCSD DDoS Attack 2007 Dataset,” Center for Applied Internet Data Analysis, University of California, San Diego, CA, USA, 2007. [Online]. Available: https://www.caida.org/catalog/datasets/ddos-20070804_dataset/
- [2] M. Alkasassbeh, G. Al-Naymat, A. B. A. Hassanat, and M. Almseidin, “Detecting distributed denial of service attacks using data mining techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, pp. 436–445, 2016, doi: 10.14569/IJACSA.2016.070159.
- [3] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in *Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP)*, Funchal, Portugal, 2018, pp. 108–116.
- [4] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, “Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy,” in *Proc. IEEE 53rd International Carnahan Conference on Security Technology (ICCST)*, Chennai, India, 2019, doi: 10.1109/CCST.2019.8888419.
- [5] Canadian Institute for Cybersecurity, “CICDDoS2019: DDoS Evaluation Dataset,” University of New Brunswick, Fredericton, NB, Canada, 2019. [Online]. Available: <https://www.unb.ca/cic/datasets/ddos-2019.html>
- [6] Canadian Institute for Cybersecurity, “CICIDS2017 Dataset,” University of New Brunswick, Fredericton, NB, Canada, 2017. [Online]. Available: <https://www.unb.ca/cic/datasets/ids-2017.html>
- [7] Canadian Institute for Cybersecurity, “CSE-CIC-IDS2018 Dataset,” University of New Brunswick, Fredericton, NB, Canada, 2018. [Online]. Available: <https://www.unb.ca/cic/datasets/ids-2018.html>
- [8] D. Erhan and E. Anarım, “Boğaziçi University distributed denial of service dataset,” *Data in Brief*, vol. 32, Art. no. 106187, 2020, doi: 10.1016/j.dib.2020.106187.
- [9] M. S. Elsayed, N.-A. Le-Khac, and A. D. Jurcut, “InSDN: A novel SDN intrusion dataset,” *IEEE Access*, vol. 8, pp. 165263–165284, 2020, doi: 10.1109/ACCESS.2020.3022633.
- [10] N. M. Yungaicela-Naula, C. Vargas-Rosales, J. A. Pérez-Díaz, E. Jacob, and C. Martinez-Cagnazzo, “Physical assessment of an SDN-based security framework for DDoS attack mitigation: Introducing the SDN-SlowRate-DDoS dataset,” *IEEE Access*, vol. 11, pp. 46820–46831, 2023, doi: 10.1109/ACCESS.2023.3274577.

- [11] A. A. Bahashwan, M. Anbar, S. Manickam, G. Issa, M. A. Aladaileh, B. A. Alabsi, and S. D. A. Rihan, "HLD-DDoSDN: High and low-rates dataset-based DDoS attacks against SDN," PLOS ONE, vol. 19, no. 2, Art. no. e0297548, 2024, doi: 10.1371/journal.pone.0297548.
- [12] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón, "UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs," Computers & Security, vol. 73, pp. 411–424, 2018, doi: 10.1016/j.cose.2017.11.004.
- [13] M. Tavallaei, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), Ottawa, ON, Canada, 2009, pp. 1–6.
- [14] S. Shiravi, H. Shiravi, M. Tavallaei, and A. A. Ghorbani, "Toward a systematic approach to generate benchmark datasets for intrusion detection," Computers & Security, vol. 31, no. 3, pp. 357–368, 2012, doi: 10.1016/j.cose.2012.01.012.
- [15] T. F. Deressu, A. M. Beyene, and A. Y. Gemechu, "A survey of application layer distributed denial of service detection: Approaches, models, and datasets," Artificial Intelligence Review, vol. 59, Art. no. 139, 2026, doi: 10.1007/s10462-026-11528-3.
- [16] W. Hill, Y. T. Acquaah, J. Mason, D. Limbrick, S. Teixeira-Poit, C. Coates, and K. Roy, "DDoS in SDN: A review of open datasets, attack vectors and mitigation strategies," Discover Applied Sciences, vol. 6, Art. no. 472, 2024, doi: 10.1007/s42452-024-06172-x.
- [17] W. Ashraf and F. S. Masoodi, "A survey of intrusion detection datasets for communication networks," Discover Computing, vol. 29, Art. no. 577, 2026, doi: 10.1007/s10791-026-10486-2.

