Detecting and Mitigating Distributed Denial of Service Attacks

Alexander Hinz

Hochschule für Angewandte Wissenschaften Hamburg
Department der Informatik
Berliner Tor 5, 20099 Hamburg, Deutschland
alexander.hinz2@haw-hamburg.de

Abstract. This paper provides an overview of Distributed Denial-of-Service (DDoS) attacks, mitigation strategies and current focus of security research on the topic. Motivated by the threat automated attacks represent, special emphasis is placed on automatic defensive strategies. The paper reviews a taxonomy of DDoS attacks and defenses, distinguishing between preventive and reactive measures, and further classifying them by deployment location and detection strategy. Preventive approaches, such as protocol hardening, disabling amplification vectors, takedown efforts, as well as anomaly-based reactive measures, are reported on along with their limitations. A review of recent research (2023-2025) underscores the diversity of attack and mitigation strategies, but also reveals that only a few proposals address automatic detection in-depth. While preventive actions are essential for protecting core infrastructure, they are insufficient on their own. Mitigating DDoS attacks requires preventive and reactive techniques. The paper then delves into machine learning model-based detection by first reporting on challenges inherent in developing such solutions, before exploring proposals implementing automatic DDoS detection using a machine learning-based distributed detection systems. The use of prefix-level clustering and monitoring proposed by one work is especially interesting for future work.

Keywords: DDoS · DoS

1 Introduction

The Internet is an integral part of the so-called critical infrastructure. Critical Infrastructure is a term to describe organizations and institutions of great importance to the general public [37]. The Federal Office for Information Security (BSI) even calls the Internet a "basic prerequisite for society as a whole" [35].

It is a network of networks connecting networks and end systems with each other. The history of the Internet began approximately 60 years ago and it is still developing along with demands from new technologies. Its main function is the transportation of data, i.e., to provide the service of packet transportation to the applications. As such the Internet can be divided into the core providing

the infrastructure and the edge where the end systems such as desktops, mobile phones or IoT devices connect to it [20]. Internet infrastructure is the deployed system of network and service components which jointly provide the worldwide Internet services. So everything from the fiber, cables and routers, as well as the Domain Name System (DNS) and routing protocols, to data centers are part of the Internet Infrastructure, whereas user-faced applications such as email are excluded from it [6, 35].

Distributed-Denial-of-Service (DDoS) attacks pose a high risk to critical infrastructure [35, 4]. During DDoS attacks an adversary aims to achieve either resource or bandwidth consumption of a target to cut it off from the rest of the Internet [3]. Furthermore, the BSI finds in its report on the state of IT-Security in Germany for the year 2024 that the quality and quantity of DDoS attacks have increased [4]. The advancement of Artificial Intelligence (AI) is likely to further increase the risk of attacks, as assessments of the National Cyber Security Centre (NCSC) [30] and its German pendant the BSI report [5]. They argue that Large-Language-Models (LLMs) at first increase the ease of malware creation and performing attacks before actors with more resources are able to use AI to considerably improve their capabilities. In both cases (actors with a lack and those with an abundance of resources), an increase in frequency and intensity of cyber threats is expected. While attackers can use AI to automate and aid evasion and scalability, the same is true for defenders, who can implement measures of their own and benefit from the usage of AI [30].

This is where AI:AutoImmune, a project aiming to support defenders in securing the Internet, which HAWs INET Group is a part of, takes on the challenge. The focus of this paper is the automatic detection and mitigation of DDoS attacks, leading to the following question:

How can DDoS attacks be automatically detected and mitigated?

To approach this topic an overview of DDoS attacks, targets and defenses is provided first, before taking a deeper look into the current automatic defensive measures being developed. The remainder of this paper is structured as follows. Chapter 2 covers the background of DDoS attacks, including the history, classification of DDoS attacks as well as DDoS mitigations and DDoS topics currently discussed in the major security conferences. Chapter 3 follows with a specific focus on the automatic detection and mitigation of attacks. Here challenges when developing machine learning (ML) models for detection are presented. Furthermore, two papers on developing such a model are examined to illustrate how these challenges manifest in practice. The results of this examination guide recommendations for further directions in the conclusion.

2 DDoS Background

The first DDoS attacks were reported in the early 2000s [15, 7, 36, 32]. While first attacks have been motivated by activism and pranksters, they were then com-

mercialized and exploited by individuals and nations alike [3, 8]. DDoS attacks are often facilitated by botnets that are hosted via specialized services namely Bulletproof Hosters (BPH). BPHs ignore abuse complaints and host booter services, which provide DDoS-for-hire platforms that are resilient to takedowns [15].

The danger they pose to the Internet remains and the damage caused by them only seems to increase [16, 36, 39]. As such countermeasures [7] and taxonomies have been developed early on [25]. While new countermeasures are developed with time [40, 28], some long-standing issues enabling attacks remain unresolved [36, 24]. This section first covers the prevalent classes of DDoS attacks before then introducing common defense and mitigation techniques. Concluding with a look into current publications on the topic of DDoS and DDoS mitigation.

2.1 Classification of DDoS Attacks

Mirkovic and Reiher [25] used a diverse set of characteristics such as the degree of automation, exploited weakness, impact on the victim and validity of the source address(es) used for the DDoS attack to classify them. From this comprehensive classification the distinction based on exploited weakness (semantic or brute-force) as well as validity of the source address(es) (spoofed and non-spoofed) has been used in following works with the addition of direct and reflected attacks [17, 15].

The difference between semantic and brute-force, also called volumetric, attacks [17], lies in the way the Denial-of-Service (DoS) is achieved. Semantic attacks use the characteristics of the attacked applications themselves to cause the DDoS [25, 17]. Whereas Mirkovic and Reiher [25] use TCP SYN flood [12] as an example of a semantic attack that consumes the resources of an application, Jonker et al. [17] explain that a semantic attack could also use a malformed packet to cause a crash of the target. A TCP SYN flood is an attack, in which the attacker sends a large number of TCP SYN packets to a server, causing it to allocate resources for half-open connections that are never completed. This can eventually exhaust the server's ability to handle new connections, disrupting legitimate access [12]. While a semantic attack may cause DoS by having the target consume its resources such as memory or processing power, a volumetric attack simply overwhelms the network link of the target, thus rendering it unreachable [12, 25, 32]. The distinction between semantic and volumetric attacks is often subtle, as both lead to a denial of service, but their mechanisms differ. Semantic attacks can be mitigated by hardening protocols or deploying defensive measures, whereas volumetric attacks cannot be patched and require other forms of mitigation [25], e.g., traffic filtering. Note that semantic attacks are target specific attacks whereas volumetric attacks are not [17]. Note also that a TCP SYN flood can become a volumetric attack if the bottleneck is moved (e.g., by implementing rate limiting) from system resources towards the bandwidth [25]. This may also explain why volumetric attacks are so prevalent in the Internet [15, 19, 28, 33].

Volumetric and semantic attacks sometimes use either spoofed or non-spoofed source addresses [25, 17, 15]. An example of a semantic attack using a non-spoofed address is the state exhaustion attack called SlowLoris, developed by Robert Hansen [9, 15]. For volumetric spoofed attacks the source addresses are often selected randomly which is why they are called randomly-spoofed DoS (RSDoS) attacks [15]. Using spoofed addresses obscures the direct-path of the attack and hinders mitigation [25, 36]. While semantic DoS attacks are usually direct [25], volumetric attacks can be further classified into direct-path and reflection attacks [17, 15]. Direct-path attacks send the attack packets directly to the target from machines controlled by the attacker (e.g., a single machine or a botnet) [17]. A botnet is a number of machines connected to the Internet that have been compromised by a malicious entity to use in attacks such as a DDoS [32].

Reflection attacks have been known for more than 20 years [32] and have continued to gain prevalence with time [33, 17, 19, 29, 28, 18]. Reflective amplification attacks are a prevalent form of Distributed Denial-of-Service (DDoS) attack in which attackers send packets to third-party servers (reflectors) that, in turn, send amplified responses to the victim, overwhelming its resources while obscuring the attack's true origin [15, 34]. This method exploits the characteristics of Internet protocols, especially UDP-based services including NTP, DNS, and LDAP that respond to small requests with much larger replies, enabling attackers to maximize the volume of attack traffic with minimal effort [34, 33]. For example, NTP can amplify traffic by a factor of up to 4670, and remains together with DNS the most popular protocol abused for such attacks [33, 28, 29, 19]. Attackers often use a random mix of multiple known and newly discovered amplifiers, with 80% of attack events involving between 10 to 100 amplifiers [29]. The advantages for attackers are the ability to disguise their identity and avoid saturating their own bandwidth while using multiple amplifiers at once [34, 33]. The ease of use and the effectiveness of reflection-amplification attacks have contributed to their prevalence in the DDoS landscape in recent years [34, 15].

2.2 Defenses Against DDoS

Chang [7] organized Distributed Denial-of-Service (DDoS) defense mechanisms into three main lines of defense: attack prevention and preemption, attack detection and filtering, and attack source traceback and identification. Mirkovic and Reiher [25] introduced a taxonomy of defensive measures based on several dimensions, including the degree of cooperation required, the deployment location within the network and the activity level of the defense (preventive, reactive). The dimension activity level of the defense maps onto Chang's [7] three lines of defense with attack prevention and attack detection and filtering (reactive) are the main categories.

Reactive measures can be divided into pattern-based approaches, which rely on signatures of known attacks, and anomaly-based methods, which identify deviations from normal behavior [25]. Further, anomaly detection itself can be

based on standard thresholds or on trained models that must be carefully maintained and updated to adapt to evolving traffic patterns [25].

While Mirkovic and Reiher [25] arrange the dimension of location of the defensive measure on the same level as its activity level, Chang [7] applies it only to the reactive measure. Chang [7] also differentiates between four locations along the network path—from the victim's own network for implementing countermeasures (directly at the victim, the victim ISP, further upstream, attack source network) whereas Mirkovic and Reiher [25] only differentiate between three (victim network, intermediate network, attack source network). Detection is generally easier at the victim, while filtering is most effective when applied as close to the attack source as possible to minimize collateral damage [7]. Therefore, Chang [7] suggested having the victim notify its ISP once it detects an attack, which is how DDoS protection services operate nowadays [16, 17, 15].

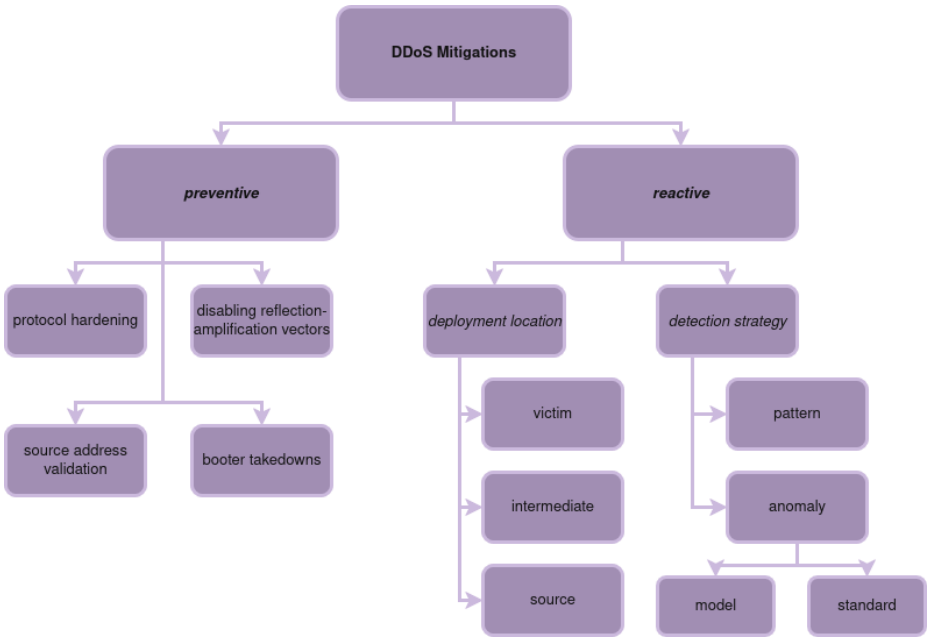

Fig. 1. Classification of DDoS mitigation tactics [7, 25]

DDoS mitigations are classified based on their activity level, deployment location and detection strategy. This provides an overview that is detailed enough without missing the forest for the trees and grounded in the presented prior work [7, 25]. Figure 1 illustrates this classification.

We first go into preventive measures and challenges in implementing them before taking a closer look at detection and mitigation methods.

2.2.1 Prevention: Among the preventive measures are protocol hardening, disabling reflection-amplification vectors, promoting the use of source address validation and booter takedowns.

Protocol hardening refers to the modification of protocols abused in attacks to prevent or at least reduce their impact [25, 33]. Measures include the introduction of rate limiting on a per-subnet basis, request-response symmetry and the introduction of session handling [33, 34]. Hardening a protocol usually has side effects including degrading protocol efficiency (e.g. through enforcing request/response symmetry) or making it vulnerable to DoS [33, 34].

Disabling reflection-amplification vectors aims at reducing the options a perpetrator has to orchestrate an attack. This can be done by either decommissioning vulnerable servers, patching vulnerabilities or disabling parts of protocols offering such vectors [15, 18, 34]. One example of the last mentioned is NTP where through disabling the command "get monlist" that is of no operational importance, an extraordinary amplification-vector can be removed [15]. For other vectors this is not as easy as identifying and disabling services with response request asymmetry takes time or is not feasible [34]. That decommissioning vulnerable servers has limited success as well [15] is another argument to follow Rossow's [33] suggestion of using secure service configurations as the default, instead of insecure ones.

Another area where the use of default configurations could reduce the prevalence of DDoS attacks is source address validation (SAV) [24]. SAV stops spoofing and therefore would completely prevent spoofing-based attacks such as reflection and RSDoS [15, 24, 32, 33]. The Spoofer measurement project from CAIDA [24] developed a way to measure whether SAV is activated in a network and argues that missing incentives are a reason for the lack of deployment. Luckie et al. further argue that internalizing the cost of not using SAV through regulation could be a way to further promote its deployment [24].

The takedown of booter services on the other hand seems to not have a lasting effect [15]. Collier et al. [8] measured the impact of takedowns and showed that there is no consistent effect. They observe a reduction in recorded attack numbers after publicized court cases and takedowns, but even the effect of wide-ranging disruptions only last between 10 to 13 weeks [8]. However, Collier et al. also report that awareness campaigns in the UK deterred potential new customers of booter services [8]. Vu et al. [38] find that while booters are resilient, safety perceptions of users and providers are changing. This suggests that takedowns do have their place in DDoS prevention, even if they only act as awareness campaigns.

2.2.2 Detection and Filtering: Detection and filtering of traffic as a measure against DDoS attacks has been either implemented as discussed by Chang [7] as a service that so-called DDoS protection services (DPS), CDNs or IXPs provide [15, 16, 19] and is being developed further [26, 40, 41]. Filtering at the victim network does allow for a fast response once an attack has been detected, but forces the victim to bear the brunt of the attack [15].

A DPS redirects traffic either by DNS or BGP into their own infrastructure to filter it [16]. The customer notifies the provider when an attack is noticed who then uses deep-packet inspection or application proxies to identify and block DDoS traffic at network or application layers [15]. This method is called scrubbing as legitimate traffic reaches the network while the "dirty" traffic is removed [19]. Jonker et al. [16] found that there has been an increase in adoption of cloud-based DPSs with a relative growth of 1.24x compared to the previous 1.5 years, which is pushed by large webhosters enabling such services for millions of addresses at once.

Another option is to discard DDoS traffic for specific prefixes in the Internet core before reaching the target network [19, 40]. This is called blackholing or remote triggered blackholing (RTBH) and leads to collateral damage as legitimate traffic from the prefix(es) identified as the attack source is dropped as well [15]. This would explain why Kopp et al. [19] found that only 3.82% of DDoS attacks are blackholed when they looked at Internet traffic captured at a major IXP. Nawrocki et al. [27] analyzed RTBH at a large European IXP finding that it only drops about 50% of unwanted traffic and that collateral damage is a minor consideration.

Ryba et al [34] provide an overview of methods for detecting and predicting amplification attacks, but also conclude that most methods focus on filtering at the victim and are attack specific. Wichtlhuber et al. [40] argue for a distributed system to be deployed at IXPs instead of approaches detecting and filtering close to the edge. They developed a system to scrub traffic which is further analyzed as an example of developing an ML model for automatic detection in section 3. The rationale for using a model-based system is its ability to adapt to changing attacker behavior [26].

A truly comprehensive solution would require cooperation between multiple ISPs and or IXP, which is challenging [15]. However, efforts to support cooperative filtering are another piece in the puzzle of DDoS mitigation [15]. Wagner et al. [39] developed a collaborative DDoS Information Exchange Point (DXP) to report amplification DDoS reflectors or targets to improve detection and mitigation. They underscore the opportunity of cooperation. Simply exchanging information between network providers makes it possible to detect and mitigate the majority of attacks, dropping as much as 90% of traffic locally [39]. Further, they listed the following challenges for detecting the attacks [39]:

- short duration of attacks (95% of attacks lasting less than 50 minutes)
- usage of multiple protocols (most attacks involve 3 or more amplification protocols)
- setting appropriate thresholds (thresholds local to one network fail to detect attacks routed via multiple reflectors)

Wagner et al. also confirmed the detection method Kopp et al. [19] used for differentiating between attack and benign traffic during a reflection-amplification attack. The filter Kopp et al. proposed is a threshold of 1 Gbps inbound traffic from more than 10 reflector IPs with the same source port [19]. This is an example of a global threshold [39].

2.3 A Look Into Current DDoS Research Landscape

A look at DDoS related papers submitted to the top four security conferences (ACM CSS, NDSS, IEEE S&P, USENIX Security) reveals 14 related papers. This signals a continued interest on the topic. The papers are subsequently classified, before taking a closer look at the proposed mitigation measures.

2.3.1 Attacks: Starting with papers focused on attacks, both semantic [1, 11, 14, 22, 45] as well as volumetric [13, 21, 31, 42] attacks are represented. Papers covering semantic attacks all report on DNS vulnerabilities leading to a DoS of the domain resolution service. Duan et al. for example developed an approach to dissect the name resolution process to study amplification vulnerabilities in DNS, uncovering compositional amplification (CAMP) vulnerabilities that can be used against the DNS infrastructure itself [11]. E.g., a CAMP attack against an arbitrary nameserver exploits several vulnerabilities to achieve a multiplication of the amplification factor. By using NS records, it triggers multiple independent queries, which cascade into further queries, ultimately consuming resources and resulting in a DoS [11]. Heftrig et al. [14] (KeyTrap) and Zhang et al. [45] (RUC) cover vulnerabilities in DNSSEC specifically.

The DNS also plays a central role as a reflector and amplifier in Li et al. [21] and Xu et al. [42] who describe techniques for volumetric DDoS attacks. While Pan et al. [31] and Guo et al. [13] also focus on volumetric attacks, they present more complex approaches (application layer traffic loops and CDN infrastructure as a reflector).

2.3.2 Mitigation: The papers solely focusing on mitigative measures are split between prevention and detection. Vu et al. [38] cover the effects of an ongoing global intervention including booter takedowns, while Yoo et al. [43] and DeLaugther and Sollins [10] both propose mechanisms extending TCP for additional shielding against misuse in attacks. Yoo et al. [43] propose using programmable switches to improve existing SYN-cookies, whereas DeLaugther and Sollins [10] introduce proof-of-work (PoW) to TCP as a highly scalable technique.

All of the papers written on DDoS attack detection are based on programmable switches, distributed, focus on scalability and use anomaly-based detection [26, 41, 46]. Zhou et al. [46] (Mew) and Wu et al. [41] (Lemon) did not directly develop methods to detect attacks but rather present improvements in measuring traffic, which in turn is then the basis for the application of detection methods (e.g., statistic-based for Lemon). Both systems claim to be resource friendly and scalable. Mew is developed specifically as an adaptive link-flooding defense system, allowing changes in defense policies without halting switches [46]. Lemon is a routing-oblivious detection system and supports flexible configurations of flow keys as well as the deployment of diverse DDoS attack detection algorithms [41]. While Zhou et al. [46] only mention Mew relying on flow state, Wu et al. [41] specify Lemon as sketch-based. A sketch is a random aggregation of IP flows, for more information on the theory behind sketches refer to Li et al. [23].

Misa et al. [26] (ZAPDOS) on the other hand use a ML model for volumetric DDoS detection, that relies on attack signatures at the source prefix level and exploit inherent clustering of addresses. Making use of a so far unused pronounced cluster-within-cluster property (Prefix-level attack signatures) to classify attack and benign source prefixes. Both attack and benign sources are not uniformly distributed across the IP address space, but instead form distinctive clusters within hierarchical prefix structures [26]. Misa et al. argue that since reflectors are typically misconfigured servers with high-bandwidth connections and stable up-time, they are mostly clustered in address regions associated with networks that have these characteristics, e.g, lax update policies and high-bandwidth connections. For botnets, Misa et al. point out that several studies confirm DDoS attacks via botnets generally do not use spoofing and that botnets accessed through a provider differ in price based on the region of the bots, thus leading to clustering in certain, more affordable regions. Further, benign traffic itself is clustered, which enables detection to focus on a small number of prefixes even when the attacker uses spoofing [26]. To hide his attack from such a detection an attacker must invest additional resources to infer source addresses closer to benign traffic [26].

Regarding the deployment location, Mew and Lemon do not specify further than "network-wide" and refer to ISPs [41, 46]. Misa et al. explicitly name the network edge as ZAPDOS intended deployment location, i.e., the victim network, as their approach is suited to networks with tight resource constraints [26].

Table 1. Overview of DDoS detection proposals

Name	**Location**	**Strategy**	**Attacks**	**Source**
Mew	intermediate	(dynamic) standard-based	volumetric (link-flooding)[1]	[46]
Lemon	intermediate	(statistic) standard-based	volumetric	[41]
Zapdos	victim	model-based	volumetric	[26]

Table 1 provides an overview of the proposed approaches classifying them according to figure 1. Both Lemon and Mew are labeled as deployable in an intermediate location between the victim and the attacker, as they do not mention their proposal being designed for use near the attack source. Since all three detect attacks relying on traffic characteristics (flow level or prefix level), they are agnostic towards the different subclasses of volumetric attacks (spoofed/non-spoofed, direct/reflective)[2]. Mew and Lemon are both standard-based in so far

[1] While they explicitly designate it as a link-flooding defense system, their main contribution is a monitoring system. As the defense policies can be changed dynamically, it should allow for general defense against volumetric attacks.

as that they rely on methods other than ML models to determine thresholds for malicious behavior.

2.4 Main Takeaways

Volumetric attacks using reflectors or botnets are the most common kind of volumetric attack [26]. Reflection attacks are especially efficient due to the amplification factor the reflector usually provides [28, 34]. Reactive measures including anomaly-based detection and filtering of attack traffic on the other hand could completely mitigate ongoing attacks [40, 41, 44].

Misa et al. [26] describe the development of a system to automatically detect volumetric DDoS attacks, which is explored further in the next section. Using prefix-level clustering to exploit the properties of reflection and botnet attacks to differentiate them from benign traffic is especially relevant as it is novel and allows the use of more complex machine learning models with reduced monitoring overhead [26].

3 Automatic Detection of DDoS

Next we discuss the development of ML-based automatic detection methods referencing Misa et al. [26] (ZAPDOS) and Wichtlhuber et al. [40] (IXPScrubber). ZAPDOS focuses on defending at the victim network, while IXPscrubber focuses on intermediary networks (at IXPs). However, first pitfalls when applying ML to the security domain [2] are introduced, so the proposed methods can subsequently be examined using this knowledge.

3.1 Challenges in Applying ML to Detection

Arp et al. [2] examined 30 papers from top-tier security conferences regarding ten common pitfalls in design, implementation and evaluation of learning-based security systems, finding that all papers suffer from at least three pitfalls. Arp et al. explain how these undermine the validity of research results and provide recommendations to rectify or reduce the impact of the pitfalls. Those pitfalls exist because ML "requires reasoning about statistical properties of data across a delicate workflow" [2]. I.e., incorrect assumptions and experimental biases impact results heavily, leading to overly optimistic assessments of the developed solution. A complete overview of issues is not possible within the scope of this work. Thus, the goal of this section is to present the pitfalls present in most papers or directly applicable to automatic attack detection and how to mitigate them.

The most common pitfalls were sampling bias and data snooping [2]. Sampling bias occurs during the data collection phase if the true underlying distribution of the input space is not represented in the sample. This can be traced back

[2] Though this only means that they can in theory detect all subclasses, not that their performance is equal for each subclass.

to the acquisition of data for the training of the model, which is challenging. Often one has to rely on synthetic data or combine multiple data sources. To mitigate this issue, one can construct different estimates of the true distribution and analyze them individually [2]. Data snooping has three expressions (test, temporal, selective) [2]. An example of selective data snooping is the removal of outliers based on statistics of the complete data set (training and test) that are usually not available at training time [2]. To prevent data snooping, one should follow the general standard of separating training, validation and test data, consider temporal dependencies when creating dataset splits and complement the experiments with recent data.

Inappropriate threat models, lab-only evaluation and inappropriate performance measures are other prevalent pitfalls, occurring at least partially in more than half of the investigated papers [2]. ML models are usually deployed in hostile environments with attackers suspecting their existence and behaving accordingly. As such, the threat models need to be precisely defined and the system validated against it. Arp et al. especially recommend focusing on white-box attacks where possible [2]. Lab evaluations are often not realistic and solely relying on them is negligent. Therefore, one should ideally deploy the system in the real world (while keeping ethical considerations in mind) or at least approximate real-world settings by accounting for typical dynamics encountered in practice [2]. Evaluation of the model should include multiple metrics as simple and one dimensional metrics may be insufficient, while complex measures could obscure the actual performance. This is especially important for detection tasks [2].

For the field of network intrusion detection Arp et al. underline the challenge of data set collection (e.g., avoiding sampling bias) and recommend using a simpler model as a baseline [2]. When evaluating an approach only comparing against mostly identical learning models is not sufficient. Comparing to traditional methods is a must to convincingly justify the overhead introduced by complex models and ML approaches in general [2]. As DDoS detection relies on classifying attack and benign traffic, label inaccuracy has to be avoided as well. To achieve this the labels need to be as accurate as possible and the remaining uncertainty has to be considered as well [2].

The other pitfalls are biased parameter selection (relating to hyperparameter and threshold selection), base rate fallacy (ignoring the base rate during evaluation) and spurious correlations (the model mistaking correlation for causation). A relevant example of spurious correlation is a model learning to detect IP ranges instead of attack patterns if most attacks originate in one network region [2]. To prevent this, explainability techniques should be applied [2]. While some pitfalls may be unavoidable, corrective measures should always be taken and resulting limitations discussed [2].

3.2 Looking at Two Development Approaches for Automatic DDoS Detection

After reporting on the challenges inherent to developing a ML-model for security and specifically detection, we now examine two examples of developing such a

model. ZAPDOS [26] and IXPScrubber [40] serve as case studies illustrating how these challenges apply detection models.

Misa et al. [26] use multiple data sources as well as synthetic data for the development of ZAPDOS. Misa et al. created their own training data by generating attack traffic and merging it with benign traffic. To approximate a realistic attack setting they use data from the Mirai botnet, booter datasets as well as a special method for creating synthetic attack sources representing spoofing-based attacks for their attack source set. This attack source set is then used to generate the attack traffic using six common attack types (reflection-amplification via DNS, NTP and flooding attacks). As a source for benign traffic Misa et al. use data from the 2019 MAWILab dataset. They aim to provide realistic attack address distributions as well as packet-level data and take extra caution to ensure proper separation of training and testing sets, using this method to create the training, testing and evaluation data. Misa et al. use false-negative and false-positive rates as their sole evaluation metrics when comparing their model to two sketch-based statistical approaches. Adversarial considerations are discussed including attacks with specific knowledge of how the model was trained.

Their proposed method could be affected by the following pitfalls: data snooping, sampling bias, spurious correlations, lab-only evaluation, inappropriate performance measures and label inaccuracy. The reason lies in their chosen data set (its creation method) and their evaluation set up. Firstly the use of 2019 MAWI data could lead to spurious correlations, data snooping, sampling bias and label inaccuracy if their generated traffic has properties diverging from it. E.g., the model may identify a pattern in the generated traffic that is irrespective of an attack and use this to differentiate the generated traffic from the MAWI traces. This would then result in the model detecting traffic similar to the pattern instead of attack traffic. While Misa et al. do acknowledge the challenge in data selection, they only mention the limited generalizability to network settings beyond MAWI.

The validity of their method could have been strengthened by using explainability techniques, recent and "natural" traffic captures (during development and evaluation), a more diverse set of measures. Furthermore Misa et al. could have discussed their labeling method as one could assume from their paper that they simply labeled their generated traffic as malicious and the MAWI traffic as benign. This would rest on the (unlikely) assumption that the MAWI trace does not contain any attack traffic.

Wichtlhuber et al. [40] underscore the importance of training data quality and the ML pipeline itself for developing functional models. They use two different data sets, the ML training set mainly used for training and a self-attack set (SAS) mainly used for validation. The training set is created using at IXPs blackholed traffic. Wichtlhuber et al. explain their labeling method and how they generate high-quality training data to mitigate sampling bias and label inaccuracy as well as consider the representativeness of their data. Wichtlhuber et al. use blackholed traffic after balancing and filtering it so that the data set has an equal distribution of blackholing flows and benign flows while maintaining

the traffic properties of the blackholed traffic [40]. The SAS data is collected by an IXP commissioning a DDoS attack on itself and represents a ground truth baseline that they use during the evaluation to evaluate the validity of their method. They explicitly address how spurious correlations and sampling bias can be reduced using this second data set. Wichtlhuber et al. further integrate local explainability (i.e., explaining model decisions for certain inputs) into their model, helping understand classification decisions during the evaluation. The separation of training, testing and evaluation data as well as adversarial security considerations are covered as well. The evaluation is carried out at five commercial IXPs using diverse metrics to assess performance (recall, precision, false positive and negative rates), investigate model drift and comparing it against a simpler baseline (a rule-tagging-based classifier), following the recommendations laid out by Arp et al. [2].

Conclusion and Outlook

DDoS attacks are a long-standing threat that continues to grow. They directly threaten the critical infrastructure of the Internet. We examined specific classes of DDoS attacks, mitigation strategies, and current challenges. Previous work found that reflection and botnet attacks make up the majority of volumetric attacks and that adversaries are adaptable. Automatic model-based detection is a measure against such adversaries. A model-based DDoS detection system should be distributed, scalable, resource efficient and incentivize cooperation. It should furthermore enable the incorporation of problem-specific knowledge such as known reflectors or cluster properties of traffic.

The quality of the training data and ML workflow are paramount to the development of the model. Preventing sampling bias and label inaccuracies during development is challenging but necessary. An appropriate baseline and multiple appropriate metrics should be used during the evaluation. For the detection of volumetric reflection attacks the threshold proposed by Kopp et al. and confirmed by Wagner et al. is such a baseline. Explainability techniques should be considered to prevent spurious correlations. Finally, the adversarial environment has to be taken into account during the design of the ML model.

Investigating clustering of attack sources at the prefix-level through the application of clustering algorithms like K-Means or hierarchical clustering is the next step we plan to take. Further, we could examine statistical properties (e.g., entropy) of attack sources and benign traffic to better understand their underlying properties. As for data sources, the CAIDA Network telescope can serve as a source for spoofed DDoS traffic and the MAWI traces can be further explored. Using scanners to periodically identify potential reflectors would allow us to be privy to changes in their prefix-level distribution.

References

[1] Yehuda Afek, Anat Bremler-Barr, and Shani Stajnrod. "NRDelegation-Attack: Complexity DDoS attack on DNS Recursive Resolvers". In: *Proceedings of the 32nd USENIX Security Symposium*. 32nd USENIX Security Symposium (USENIX Security 23): USENIX Association, 2023. ISBN: 978-1-939133-37-3. https://www.usenix.org/conference/usenixsecurity23/presentation/afek.

[2] Daniel Arp et al. "Dos and Don'ts of Machine Learning in Computer Security". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 3971–3988. ISBN: 978-1-939133-31-1. https://www.usenix.org/conference/usenixsecurity22/presentation/arp.

[3] Richard R. Brooks et al. "Distributed Denial of Service (DDoS): A History". In: *IEEE Annals of the History of Computing* 44.2 (Apr. 2022), pp. 44–54. ISSN: 1058-6180, 1934-1547. DOI: 10.1109/MAHC.2021.3072582. https://ieeexplore.ieee.org/document/9404833/.

[4] BSI. *Die Lage der IT-Sicherheit in Deutschland 2024*. Tech. rep. Bundesamt für Sicherheit in der Informationstechnik, 2024, p. 114. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2024.html?nn=129410.

[5] BSI. *Einfluss von KI auf die Cyberbedrohungslandschaft*. Tech. rep. Bundesamt für Sicherheit in der Informationstechnik, 2024, p. 11. https://www.bsi.bund.de/DE/Service-Navi/Presse/Pressemitteilungen/Presse2024/240430_Paper_Einfluss_KI_Cyberbedrohungslage.html.

[6] Vinton G. Cerf. "Preserving the internet". In: *Communications of the ACM* 65.4 (Apr. 2022), pp. 5–5. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3522782. https://dl.acm.org/doi/10.1145/3522782.

[7] R.K.C. Chang. "Defending against flooding-based distributed denial-of-service attacks: a tutorial". In: *IEEE Communications Magazine* 40.10 (Oct. 2002), pp. 42–51. ISSN: 0163-6804. DOI: 10.1109/MCOM.2002.1039856. http://ieeexplore.ieee.org/document/1039856/.

[8] Ben Collier et al. "Booting the Booters: Evaluating the Effects of Police Interventions in the Market for Denial-of-Service Attacks". In: *Proceedings of the Internet Measurement Conference*. Amsterdam Netherlands: ACM, Oct. 2019, pp. 50–64. ISBN: 978-1-4503-6948-0. DOI: 10.1145/3355369.3355592. https://dl.acm.org/doi/10.1145/3355369.3355592.

[9] Evan Damon et al. "Hands-on denial of service lab exercises using SlowLoris and RUDY". In: *Proceedings of the 2012 Information Security Curriculum Development Conference*. Kennesaw Georgia: ACM, Oct. 2012, pp. 21–29. ISBN: 978-1-4503-1538-8. DOI: 10.1145/2390317.2390321. https://dl.acm.org/doi/10.1145/2390317.2390321.

[10] Samuel DeLaughter and Karen Sollins. "SYN Proof-of- Work: Improving Volumetric DoS Resilience in TCP". In: *2025 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, May 2025, pp. 1877–

1890. ISBN: 979-8-3315-2236-0. DOI: 10.1109/SP61157.2025.00166. https://ieeexplore.ieee.org/document/11023284/.

[11] Huayi Duan et al. "CAMP: Compositional Amplification Attacks against DNS". In: *Proceedings of the 33rd USENIX Security Symposium*. Philadelphia, PA, USA: USENIX Association, 2024. ISBN: 978-1-939133-44-1. https://www.usenix.org/conference/usenixsecurity24/presentation/duan.

[12] W. Eddy. *TCP SYN Flooding Attacks and Common Mitigations*. Tech. rep. RFC4987. RFC Editor, Aug. 2007, RFC4987. DOI: 10.17487/rfc4987. https://www.rfc-editor.org/info/rfc4987.

[13] Run Guo, Baojun Liu, and Haixin Duan. "Temporal CDN-Convex Lens: A CDN-Assisted Practical Pulsing DDoS Attack". In: *Proceedings of the 32nd USENIX Security Symposium*. Anaheim, CA, USA: USENIX Association, 2023. ISBN: 978-1-939133-37-3. https://www.usenix.org/conference/usenixsecurity23/presentation/guo-run.

[14] Elias Heftrig et al. "The Harder You Try, The Harder You Fail: The KeyTrap Denial-of-Service Algorithmic Complexity Attacks on DNSSEC". In: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. Salt Lake City UT USA: ACM, Dec. 2024, pp. 497–510. ISBN: 979-8-4007-0636-3. DOI: 10.1145/3658644.3670389. https://dl.acm.org/doi/10.1145/3658644.3670389.

[15] Raphael Hiesgen et al. "The Age of DDoScovery: An Empirical Comparison of Industry and Academic DDoS Assessments". In: *Proceedings of the 2024 ACM on Internet Measurement Conference*. Madrid Spain: ACM, Nov. 2024, pp. 259–279. ISBN: 979-8-4007-0592-2. DOI: 10.1145/3646547.3688451. https://dl.acm.org/doi/10.1145/3646547.3688451.

[16] Mattijs Jonker et al. "Measuring the Adoption of DDoS Protection Services". In: *Proceedings of the 2016 Internet Measurement Conference*. Santa Monica California USA: ACM, Nov. 2016, pp. 279–285. ISBN: 978-1-4503-4526-2. DOI: 10.1145/2987443.2987487. https://dl.acm.org/doi/10.1145/2987443.2987487.

[17] Mattijs Jonker et al. "Millions of targets under attack: a macroscopic characterization of the DoS ecosystem". In: *Proceedings of the 2017 Internet Measurement Conference*. London United Kingdom: ACM, Nov. 2017, pp. 100–113. ISBN: 978-1-4503-5118-8. DOI: 10.1145/3131365.3131383. https://dl.acm.org/doi/10.1145/3131365.3131383.

[18] Maynard Koch et al. "Forward to Hell? On the Potentials of Misusing Transparent DNS Forwarders in Reflective Amplification Attacks". In: *Proceedings of ACM Conference on Computer and Communications Security (CCS)*. New York: ACM, 2025. DOI: https://doi.org/10.48550/arXiv.2510.18572.

[19] Daniel Kopp, Christoph Dietzel, and Oliver Hohlfeld. "DDoS Never Dies? An IXP Perspective on DDoS Amplification Attacks". In: *Passive and Active Measurement: 22nd International Conference, PAM 2021, Virtual Event, March 29 – April 1, 2021, Proceedings*. Cottbus, Germany: Springer-

Verlag, 2021, pp. 284–301. ISBN: 978-3-030-72581-5. DOI: 10.1007/978-3-030-72582-2_17. https://doi.org/10.1007/978-3-030-72582-2_17.

[20] James F. Kurose and Keith W. Ross. *Computer networking: a top-down approach*. Eighth edition. OCLC: 1124771875. Hoboken, NJ: Pearson, 2021. ISBN: 978-0-13-668155-7.

[21] Xiang Li et al. "DNSBomb: A New Practical-and-Powerful Pulsing DoS Attack Exploiting DNS Queries-and-Responses". In: *2024 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, May 2024, pp. 4478–4496. ISBN: 979-8-3503-3130-1. DOI: 10.1109/SP54263.2024.00264. https://ieeexplore.ieee.org/document/10646654/.

[22] Xiang Li et al. "TuDoor Attack: Systematically Exploring and Exploiting Logic Vulnerabilities in DNS Response Pre-processing with Malformed Packets". In: *2024 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, May 2024, pp. 4459–4477. ISBN: 979-8-3503-3130-1. DOI: 10.1109/SP54263.2024.00172. https://ieeexplore.ieee.org/document/10646751/.

[23] Xin Li et al. "Detection and identification of network anomalies using sketch subspaces". In: *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. Rio de Janeriro Brazil: ACM, Oct. 2006, pp. 147–152. ISBN: 978-1-59593-561-8. DOI: 10.1145/1177080.1177099. https://dl.acm.org/doi/10.1145/1177080.1177099.

[24] Matthew Luckie et al. "Network Hygiene, Incentives, and Regulation: Deployment of Source Address Validation in the Internet". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. London United Kingdom: ACM, Nov. 2019, pp. 465–480. ISBN: 978-1-4503-6747-9. DOI: 10.1145/3319535.3354232. https://dl.acm.org/doi/10.1145/3319535.3354232.

[25] Jelena Mirkovic and Peter Reiher. "A taxonomy of DDoS attack and DDoS defense mechanisms". In: *ACM SIGCOMM Computer Communication Review* 34.2 (Apr. 2004), pp. 39–53. ISSN: 0146-4833. DOI: 10.1145/997150.997156. https://dl.acm.org/doi/10.1145/997150.997156.

[26] Chris Misa et al. "Leveraging Prefix Structure to Detect Volumetric DDoS Attack Signatures with Programmable Switches". In: *2024 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, May 2024, pp. 4535–4553. ISBN: 979-8-3503-3130-1. DOI: 10.1109/SP54263.2024.00267. https://ieeexplore.ieee.org/document/10646832/.

[27] Marcin Nawrocki et al. "Down the Black Hole: Dismantling Operational Practices of BGP Blackholing at IXPs". In: *Proceedings of the Internet Measurement Conference*. Amsterdam Netherlands: ACM, Oct. 2019, pp. 435–448. ISBN: 978-1-4503-6948-0. DOI: 10.1145/3355369.3355593. https://dl.acm.org/doi/10.1145/3355369.3355593.

[28] Marcin Nawrocki et al. "SoK: A Data-driven View on Methods to Detect Reflective Amplification DDoS Attacks Using Honeypots". In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. Delft, Netherlands: IEEE, July 2023, pp. 576–591. ISBN: 978-1-6654-6512-0. DOI:

10.1109/EuroSP57164.2023.00041. https://ieeexplore.ieee.org/document/10190540/.

[29] Marcin Nawrocki et al. "The far side of DNS amplification: tracing the DDoS attack ecosystem from the internet core". In: *Proceedings of the 21st ACM Internet Measurement Conference*. Virtual Event: ACM, Nov. 2021, pp. 419–434. ISBN: 978-1-4503-9129-0. DOI: 10.1145/3487552.3487835. https://dl.acm.org/doi/10.1145/3487552.3487835.

[30] NCSC. *Impact of AI on cyber threat from now to 2027*. Tech. rep. National Cyber Security Centre, 2025, p. 7. https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027.

[31] Yepeng Pan, Anna Ascheman, and Christian Rossow. "Loopy Hell(ow): Infinite Traffic Loops at the Application Layer". In: *Proceedings of the 33nd USENIX Security Symposium*. Philadelphia, PA, USA: USENIX Association, 2024. ISBN: 978-1-939133-44-1. https://www.usenix.org/conference/usenixsecurity24/presentation/pan-yepeng.

[32] Vern Paxson. "An analysis of using reflectors for distributed denial-of-service attacks". In: *ACM SIGCOMM Computer Communication Review* 31.3 (July 2001), pp. 38–47. ISSN: 0146-4833. DOI: 10.1145/505659.505664. https://dl.acm.org/doi/10.1145/505659.505664.

[33] Christian Rossow. "Amplification Hell: Revisiting Network Protocols for DDoS Abuse". In: *Proceedings 2014 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society, 2014. ISBN: 978-1-891562-35-8. DOI: 10.14722/ndss.2014.23233.

[34] Fabrice J. Ryba et al. *Amplification and DRDoS Attack Defense – A Survey and New Perspectives*. Version Number: 3. 2015. DOI: 10.48550/ARXIV.1505.07892. https://arxiv.org/abs/1505.07892.

[35] Johann Schlamp, Matthias Wählisch, and Thomas C. Schmidt. *Zweite Internet Backbone-Studie: Auslandsverbindungen und CDN-Kompetenz*. Abschlussbericht Projekt 415 Los 1. Bundesamt für Sicherheit in der Informationstechnik, Feb. 2022, p. 341. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Studien/ZwiBACK/ZwiBACK-Studie.pdf.

[36] Daniel Senie and Paul Ferguson. *Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing*. Request for Comments RFC 2827. Num Pages: 10. Internet Engineering Task Force, May 2000. DOI: 10.17487/RFC2827. https://datatracker.ietf.org/doc/rfc2827.

[37] *Verordnung zur Bestimmung Kritischer Infrastrukturen nach dem BSI-Gesetz*. 2016. https://www.gesetze-im-internet.de/bsi-kritisv/BJNR095800016.html (visited on 09/29/2025).

[38] Anh V. Vu et al. "Assessing the aftermath: the effects of a global takedown against DDoS-for-hire services". In: *Proceedings of the 34th USENIX Conference on Security Symposium*. SEC '25. Seattle, WA, USA: USENIX Association, 2025. ISBN: 978-1-939133-52-6. DOI: https://www.usenix.org/conference/usenixsecurity25/presentation/vu.

[39] Daniel Wagner et al. "United We Stand: Collaborative Detection and Mitigation of Amplification DDoS Attacks at Scale". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. Virtual Event Republic of Korea: ACM, Nov. 2021, pp. 970–987. ISBN: 978-1-4503-8454-4. DOI: 10.1145/3460120.3485385. https://dl.acm.org/doi/10.1145/3460120.3485385.

[40] Matthias Wichtlhuber et al. "IXP scrubber: learning from blackholing traffic for ML-driven DDoS detection at scale". In: *Proceedings of the ACM SIGCOMM 2022 Conference*. Amsterdam Netherlands: ACM, Aug. 2022, pp. 707–722. ISBN: 978-1-4503-9420-8. DOI: 10.1145/3544216.3544268. https://dl.acm.org/doi/10.1145/3544216.3544268.

[41] Wenhao Wu et al. "Lemon: network-wide DDoS detection with routing-oblivious per-flow measurement". In: *Proceedings of the 34th USENIX Conference on Security Symposium*. SEC '25. event-place: Seattle, WA, USA. Seattle, WA, USA: USENIX Association, 2025. ISBN: 978-1-939133-52-6. DOI: https://dl.acm.org/doi/10.5555/3766078.3766262.

[42] Wei Xu et al. "TsuKing: Coordinating DNS Resolvers and Queries into Potent DoS Amplifiers". In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. Copenhagen Denmark: ACM, Nov. 2023, pp. 311–325. ISBN: 979-8-4007-0050-7. DOI: 10.1145/3576915.3616668. https://dl.acm.org/doi/10.1145/3576915.3616668.

[43] Sophia Yoo, Xiaoqi Chen, and Jennifer Rexford. "SmartCookie: Blocking Large-Scale SYN Floods with a Split-Proxy Defense on Programmable Data Planes". In: *Proceedings of the 33rd USENIX Security Symposium*. Philadelphia, PA, USA: USENIX Association, 2024. ISBN: 978-1-939133-44-1. https://www.usenix.org/conference/usenixsecurity24/presentation/yoo.

[44] Adrian Zapletal and Fernando Kuipers. "Slowdown as a Metric for Congestion Control Fairness". In: *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. Cambridge MA USA: ACM, Nov. 2023, pp. 205–212. ISBN: 979-8-4007-0415-4. DOI: 10.1145/3626111.3628185. https://dl.acm.org/doi/10.1145/3626111.3628185.

[45] Shuhan Zhang, Dan Li, and Baojun Liu. "Your Shield is My Sword: A Persistent Denial-of-Service Attack via the Reuse of Unvalidated Caches in DNSSEC Validation". In: *Proceedings of the 34th USENIX Security Symposium*. Seattle, WA, USA: USENIX Association, 2025. ISBN: 978-1-939133-52-6. https://www.usenix.org/system/files/usenixsecurity25-zhang-shuhan.pdf.

[46] Huancheng Zhou et al. "Mew: Enabling Large-Scale and Dynamic Link-Flooding Defenses on Programmable Switches". In: *2023 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, May 2023, pp. 3178–3192. ISBN: 978-1-6654-9336-9. DOI: 10.1109/SP46215.2023.10179404. https://ieeexplore.ieee.org/document/10179404/.