Evading stepping stone detection by using enough chaff perturbations

No Author Given

No Institute Given

Abstract. Stepping-stones are used extensively by attackers to hide their identify and access restricted targets. Many methods have been proposed to detect stepping-stones and resist evasive behaviour, but so far no benchmark dataset exists to provide a fair comparison of detection rates. We propose a comprehensive framework to simulate realistic stepping-stone behaviour with effective evasion tools, and release a large dataset to evaluate stepping-stone detection methods. We use this data to evaluate detection rates of eight state-of-the-art methods and highlight their respective strengths and weaknesses. Our results show that currently, no method is capable to reliably detect stepping-stone when the attacker inserts suitable chaff perturbations, disproving several robustness claims.

1 Introduction

The problem of stepping-stones detection (SSD) has been studied for over 20 years, yet the body of literature fails at providing an informative overview of the detection capabilities of current methods. In this paper, we set out to do just that by evaluating and comparing a number of selected state-of-the-art approaches on a new and independently generated dataset.

In a stepping-stone attack, malicious commands are relayed via a chain of compromised hosts, called stepping-stones, in order to access restricted resources and reduce the chance of being traced back. Real-world attacks using stepping-stone chains include Operation Aurora [25], Operation Night Dragon [2], the Black Energy [19] attack on the Ukrainian powergrid, and the MEDJACK [6] attack where medical devices were used as stepping-stones. The infamous Archimedes tool [3] used by the CIA leverages stepping-stone chains to reach the LAN of target hosts, while the European Union Agency for Cybersecurity classifies stepping-stone attacks as one of the top ten threats to IoT-devices [13].

The detection of interactive stepping-stones is challenging due to various reasons. Attackers are not constrained to specific proxy techniques and can obfuscate relayed traffic with evasiive tactics. Packet-based methods are computationally expensive, with scalability and false-positives being a problem for large networks. Like many intrusion attacks, stepping-stones are rare and there exists no public data representing real stepping-stone behaviour, and researchers have to rely on synthetic data. Some attempts have been made to create publicly available stepping-stone testbeds, yet most researchers evaluate their SSD methods on self-provided private data, which makes a direct comparison of the achieved results impossible.

In this work, we provide the following contributions:

1. We describe a framework to generate data that represents realistic stepping-stone data without bias to particular detection mechanisms. Our framework is scalable and capable of generating sufficient variety in terms of network settings and conducted activity.
2. We release a large and comprehensive dataset suitable for the training of machine-learning-based methods and in-depth performance evaluation. To our knowledge, this is the first public SSD dataset.
3. We re-implemented eight SSD methods that represent the current state-of-the-art and provide a fair evaluation of their capabilities in a number of settings.
4. Our evaluation shows that while most methods can accurately detect command propagation, detection rates plummet when appropriate chaff is inserted. This result disproves the claims made for multiple methods that their detection rates are robust against chaff perturbations.

The rest of the paper is organised as following: Section 1 provides an introduction and background to the problem of stepping-stone detection. Section 2 discusses the particular design of the data generation framework. Section 3 presents the dataset arrangement in terms of background and attack data and discusses evaluation methods. Section 4 discusses the selection process, properties, and implementation of the eight SSD methods that we implemented for evaluation. Section 5 discusses the results achieved by the implemented methods on the given data. Section 6 discusses related work.

1.1 Background

Stepping-stones were first conceptualised by Staniford-Chen and Heberlein in 1995 [24]. In an interactive stepping-stone attack, an attacker located at the origin host, called *host O*, sends commands to and awaits their response from a target, *host T*. The commands and responses are proxied via a chain of one or more intermediary stepping-stone hosts, called *host* $S_1, \ldots, S_N$, such as depicted in Fig. 1. Once a host S_i is brought under control, it can be turned into a stepping-stone with simple tools and steps. Some of the most common set-ups are port forwarding via SSH-tunnels, setting up a backpipe with NetCat, or using metasploit to set up a SOCKS proxy [14].

Typically, host O is located outside the targeted network, and the attack is started by compromising an initial foothold S_1 inside the network, from wich the attacker tries to move deeper into the network to reach the most valuable target T [1]. The attacker may use one or more compromised intermediary hosts $S_1, \ldots, S_{i-1}$ outside the target network at different geographical locations in order to make tracing back the attack impossible.

Stepping-stone detection (SSD) is a process of observing all incoming and outgoing connections on a particular host h_i and determining whether it is used to relay commands. This is generally done with no prior information about any other stepping-stone hosts $S_1, \ldots S_N$ or the endpoints O and T. Once h_i is known to be a stepping-stone host, further members of $\{S_1, \ldots S_N\}$ can be traced back from h_i.

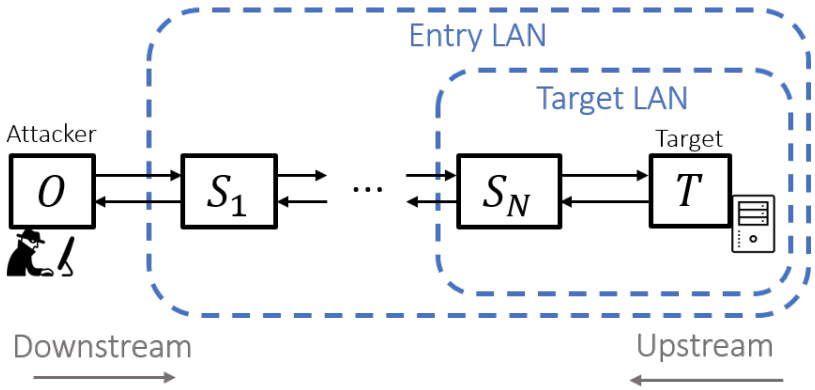

Fig. 1. Depiction of an exemplary stepping-stone chain.

A popular approach to SSD is to compare connections pairwise to identify whether they carry the same information. A wide-spread way to achieve this is via *watermarking*, an active SSD mechanism. A watermark is a unique binary string, which is usually embedded into a connection by altering the packet interarrival times. Popular passive SSD methods are based on *packet-correlation* and attempt to identify packets that appear in both connections. Since connections can be encrypted, this is often done by comparing sequences of interarrival times, packet sizes, and the overall number of packets in each connection. Anomaly-based methods aim to detect the insertion of time delays and chaff perturbations in a connection as deviations from typical TCP-behaviour to indicate of suspicious behaviour.

Another prominent approach to detect stepping stones is by measuring *round-trip-times* (RTTs). The RTT of a connection is the time it takes for a packet to be sent to the receiver plus the time it takes for an acknowledgement of that packet to be received. Relaying packets affects the overall RTT.

To avoid detection, several evasive flow transformation techniques exist that aim at decreasing observable correlation between two connections in a chain.

- **Packet transfer delays/drops**: An attacker can choose to apply artificial delays to forwarded packets, or drop certain packets to cause retransmission, in order to create temporal disparity between connections. Researchers often assume the existence of a maximum tolerable delay [12].
- **Chaff perturbations**: Chaff packets do not contain meaningful content and are added to individual connections in a chain without being forwarded. Adding chaff perturbations can be used to shape the connection profile towards other traffic types.
- **Repacketisation**: Repacketisation is the practice of combining closely adjacent packets into a larger packet, splitting a packet into multiple smaller packets, or altering the packet content to change observed packet sizes and numbers.

– **Flow splitting/merging**: An attacker can split the flow of packets to two or more connections and merge them at the target to alter the amount of packets in each connection.

In our evaluation, we set out to understand the effect of different evasive methods on detection rates.

2 Data generation setting

Our goal is to simulate data that reflects the different aspects of interactive stepping-stone behaviour in a reproducible manner. For a fair and thorough evaluation, we want to cover different settings and interactions to highlight strengths and weaknesses of different SSD methods.

2.1 Containerisation

To ensure reproducibility, we rely on containerisation. A container is a standard unit of software that runs standalone in an isolated user space in order to remove platform dependencies. Compared to virtual machines, containers always start from an identical state and are highly specialized in their purpose, with each container running only a specific piece of software.

The advantages of containers over virtual machines enable us to easily control, modify, repeat, and scale a network of containers while emulating different network settings. The use of containerisation for this project follows a traffic generation paradigm designed for machine learning, introduced by Clausen et al. [7].

2.2 Simulating stepping stones with SSH-tunnels and Docker

We want to capture data not only from one interaction in a fixed stepping-stone chain, but from many interactions and chains with different settings. For that, we run multiple simulations, with each simulation establishing a stepping-stone chain and controlling the interactions between host O and host T.

A simulation begins with the start-up of the necessary containers and ends with their takedown. We simulate host O, host T, and host $S_1, \ldots, S_n$ with SSH-daemon containers. To establish a connection chain, we connect these containers via SSH-tunnels, with the first tunnel forwarding a port from host O to host S_1, which is then forwarded to host S_2 by the second tunnel etc. As mentioned by Gordon Fraser [14], this is one of the most common pivoting methods for attackers. Traffic is captured both at host T and host S_n, which acts as the final stepping-stone in the chain. Fig. 2 depicts a packet transfer via an exemplary chain.

.

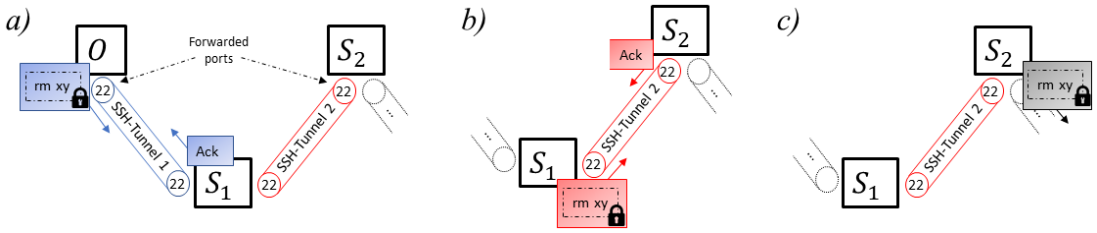

Fig. 2. Depiction of the way a command is packetised, encrypted, and travels through the different stages of the stepping-stone chain via SSH-tunnels.

Simulating interactive SSH-traffic In order to generate enough data instances representing interactive stepping stone behaviour, we automatised the communication between host O and host T. For each simulation, we generate a script which passes SSH-commands from host O to host T, with several measures introduced to taken to make them realistic.

First, each session tries to mimic a real user's action. We compiled a command database which consists of common commands and their usage frequency, similar to [31].Commands are drawn randomly according to their usage frequency and concatenated to a script. Commands can either be atomic, such as "ls-la" or "pwd", or compound. Compound commands need additional input such as the directory and name of a specific file that is transferred, or input text to fill a file. The content and sometimes length of these inputs as well as transferred files are randomised appropriately when a compound command is drawn. Scripts are of varying length and end once the *End*-command is drawn at random from the command catalogue.

To simulate human behaviour that is reacting to the response from host T, all commands are separated by *sleep*-commands for time t, which is drawn from a truncated Pareto-distribution. Paxson et al. [22] have shown that interpacket spacings corresponding to typing and "think time" pauses are well described by Pareto distributions with a shape parameter $\alpha \approx 1.0$. We use a truncated distribution capped at 10s to avoid infinite waiting times since we already included a mechanism to end a script.

Simulating different network settings Hosts in a stepping-stone chains can be separated by varying distances. Some may sit in the same LAN, while others may communicate via the Internet from distant geographical locations, which influences the round-trip-time, bandwidth, and network reliability.

Docker communication takes place over virtual bridge networks, so the throughput is far higher and more reliable than in real-world networks. To retard the quality of the Docker network to realistic levels, we rely on the emulation tool Netem, which allows users to artificially simulate network conditions such as high latency, low bandwidth, or packet corruption/drop [16].

We apply Netem commands to the network interface of each container, which adds correlated delays to incoming and outgoing packets that are drawn from a normal

distribution with mean μ, variance σ^2, and correlation ρ_1. We furthermore apply correlated packet loss and corruption drawn from a binomial distribution with probability p and correlation ρ_2. Lastly, we apply an overall limit B on the bandwidth of container network interfaces.

To allow for different types of host separation, we set the network settings and bandwidth limit for each host container individually before each simulation, and draw each of the given parameters from a suitable distribution. This allows for some hosts to experience very fast and reliable communication while others experience more congested communication. We store the drawn parameters along with the collected traffic for each simulation to include the effect of network congestion in the evaluation.

2.3 Evasive tactics

Adding transfer delays To simulate evasive behaviour, we add transfer delays to forwarded packets. This method, often called *jittering*, can destroy time-based watermarks in packet flows and help decrease observable correlation between two connections. The delays are added using NetEm. We draw delays from a uniform distribution, covering the interval $[0, \delta_D]$. This particular choice has been suggested by Padhye et al. [21] in order to mimic the interarrival distributions of streaming services. The value of δ_D is fixed before each simulation and can be varied to allow for different degrees of packet jittering.

As pointed out by Donoho et al. [12], excessively long delays often lead to difficulties in the TCP-protocol due to the significant increase of packet reordering and response time-outs. We explore values for δ_D up to 1500 ms, with values above leading to unstable communication. Results in Section 5 show that this is enough to render watermarking methods and most flow correlation methods obsolete.

Adding chaff perturbation We insert chaff packets to individual connections in the chain using a Netcat client. The inserted chaff packets do not contain actual information and act as noise to decorrelate individual connections in the chain. To add and filter packets in a connection, we open additional ports in each SSH-tunnel that are however not forwarded through the entire chain.

Again, Padhye et al. [21] suggest to generate chaff in a way that mimics the flow characteristics of streaming services in order to both spread the added perturbations evenly across the connection and increase the difficulty of detecting the perturbation itself. For this, packet sizes Are drawn from a truncated Lognormal-distribution with mean μ_C, while transmission intervals are drawn from a uniform distribution that covers the interval $[\delta_c/2, \delta_c]$ to mimic a relatively constant packet flow. By adjusting δ_C, we can control the amount of chaff sent.

Repacketisation and Flow splitting/merging By design, SSH-tunnels perform repacketisation along with re-encryption and independent packet confirmations. Flow splitting is currently not addressed by existing methods, which is why we did not include this tactic in our implementation.

Fig. 3 depicts the interplay of the different containers in our simulation.

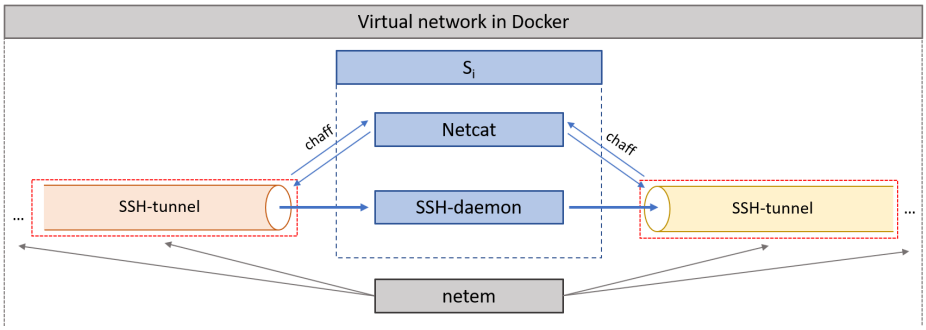

Fig. 3. Depiction the simulation setup for each host in the chain.

3 Evaluation data

We want to look at a variety of attack scenarios to highlight the strengths and weaknesses of different SSD approaches. We created three main attack datasets that contain different forms and amounts of evasive behaviour, and a smaller dataset to highlight the influence of different chain lengths.

To present a valuable false positive test, we provide three datasets with benign background traffic. The first contains general real-world traffic, while the second and third contain benign data that bears similar traffic characteristics as the generated attack data.

3.1 Stepping-stone data

We generate stepping-stone data using the setup described in Section 2. We create our main datasets using a chain of four stepping-stones S_1, S_2, S_3, and S_4. We subdivide into three datasets: We first capture data without transfer delays and chaff perturbations in **dataset BA** (**b**aseline **a**ttack). We then capture data once with added transfer delays with varying δ_D to control delays in **dataset DA** (**d**elay **a**ttack), and once with added chaff perturbations of varying δ_C in **dataset CA** (**c**haff **a**ttack). Each dataset contains 30.000 connection pairs.

We furthermore create a smaller **dataset CL** (**c**hain **l**ength) with differing numbers of stepping-stones (1,3,5, and 8 jumps) without transfer delays and chaff perturbations to evaluate this effect on RTT-based methods. For every jump, we generated 1.000 connection pairs.

For all datasets, we randomise network settings. In Section B, we examine the effect of extreme network settings on detection rates.

3.2 Benign data

We need to provide a realistic background of benign data that reflects the heterogeneous nature of regular network traffic, both for evaluation of false positives and for the training of ML-based methods.

We include real-world traffic traces, taken from the **CAIDA** *2018 Anonymized Internet Traces* dataset [4]. This data contains traces collected from high-speed monitors on a commercial backbone link, and is often used for research on the characteristics of Internet traffic, including application breakdown, security events, geographic and topological distribution, flow volume and duration.

The CAIDA dataset provides suitable general background traffic. To sufficiently test for false-positive, we also need to include benign traffic that has similar characteristics to the attack traffic and was generated in a similar network environment. We created a set of interactive SSH-connections that communicate directly between the client and the server without a stepping-stone. We follow the same procedure as described in Section 2.2 with the same randomised network congestion settings as described in 2.2. We then proceed to pair connections by random.

Since we generate perturbations with multimedia streams characteristics, we additionally want to include traffic actual multimedia stream traffic to test for false-positives. For that, we captured traffic from a Nginx-server streaming video to a client. Video content is generated randomly from a set of video files and multiple values for the streaming quality are used.

	Label	Nr. of conn.	Purpose
SS data	set BA	30,000	Baseline attack data without evasion tactics
	set DA	30,000	Inclusion of delays with varying δ_D
	set CA	30,000	Inclusion of chaff with varying δ_C
	set CL	40,000	Data from chains of different lengths, no evasion tactics
Benign data	CAIDA	60,000	General background data
	SSH	20,000	Background data similar to attack commands
	Multim.	20,000	Background data similar to chaff perturbations

Table 1. Summary of different components in our evaluation data.

We merge the three datasets to create our benign background dataset, with the CAIDA part containing 60.000 connection pairs, while the other two each contain 20.000 connection pairs. The amount of SSH traffic and multimedia streams in this setting is inflated from a realistic setting in order to highlight the strengths and drawbacks of SSD methods. Nevertheless, SSH and especially video streaming represent a significant proportion of overall network traffic (up to 0.2% of flows for SSH and up 3% for video streaming according to Velan et al. in 2016 [26]), and need to generate low false-positives

for an SSD-method to be operational. In Section .1, we analyse false-positives for each dataset individually. Table 1 summarises the different parts in our evaluation data.

3.3 Data preparation

We collect packets to connections according to the usual 5-tuple consisting of {Src. IP, Dst. IP, Src. port, Dst. port, IP protocol}. We select and pair TCP-connections at random to make connection pairs.

To create a fair playing field for the selected SSD methods, we only look at connections that exchange more than 1500 packets and exclude shorter connections from both the background data. This number seems like a suitable minimal limit for a successful interactive stepping-stone chain as all of the selected methods are supposed to make successful detection with less packets, and there were no connections with less packets in the stepping-stone dataset.

The initialisation and takedown of the SSH-tunnels usually follows a distinct pattern that can be learned and consequently boost detection rates for ML-based methods. Since this pattern is not necessarily representative of actual stepping-stone behaviour, we remove the first and last thirty packets in the attack connections.

True stepping stone connections are rare compared to benign ones, making their detection an imbalanced classification problem. An appropriate evaluation measure for imbalanced data are false positive and false negative rates as well as the *Area-under-ROC-curve* (AUC) for threshold-based methods.

4 Selected SSD methods and Implementation

A range of underlying techniques exist for SSD, and we try to include approaches from every area to create an informative overview and highlight strengths and weaknesses.

We surveyed publications to create a collection of SSD methods. We started with the publications from surveys [23, 27], and then added impactful recent publications found via Google Scholar[1].

From here, we selected approaches based on the following criteria:

1. The achieved detection and false positive rates claimed by the authors,
2. and whether the model design shows robustness against any evasion tactics as claimed by the authors.
3. We always selected the latest versions if a method has been improved or updated by the authors.

Below, we describe the selected methods. Table 2 contains a summary of the included methods. We labelled each method to make referring to it in the evaluation easier.

[1] keywords “connection”, “correlation” “stepping-stone”, “detection”, “attack”, “chaff perturbation”

Category	Approach	TP	FP	Robustness	Label
Packet-corr.	Yang, 2011 [33]	100%	0%	jitter/< 80% chaff	PContext
Neural networks	Nasr, 2018 [20]	90%	0.0002%	small jitter	DeepCorr
	Wu, 2010 [30]	100%	0%	-	WuNeur
RTT-based	Yang, 2015 [34]	not provided		50% chaff	RWalk
	Huang, 2016 [18]	85%	5%	-	Crossover
Anomaly-based	Crescenzo, 2011 [9]	99%	1%	jitter/chaff	Ano1
	Huang, 2011 [17, 11]	95%	0%	> 25% chaff/ > 0.2s jitter	Ano2
Watermarking	Wang, 2011 [28]	100%	0.5%	< 1.4s jitter	WM

Table 2. Summary of included SSD-methods along with the claimed true positive and false positive rates and evasion robustness by the corresponding authors. We added labels to each method for later reference.

PContext, 2011 Yang et al. [33] compare sequences of interarrival times in connection pairs to detect potential stepping-stone behaviour. For that, the context of a packet is defined as the packet interarrival times around that packet. The respective contextual distance between packets is estimated using the Pearson correlation. The authors focus on *Echo*-packets instead of *Send*-packets to resist evasion tactics.The authors evaluate their results with up to 100% chaff ratio with 100% detection rate.

WuNeur, 2010 A notable initial example of neural network applications to SSD came from Wu et al. [29], and which was later improved by the authors [29]. The designed neural network model is based on sequences of RTTs, which are fed into a feed-forward network to predict the downstream length of the chain. The network itself only contains one hidden layer and achieves good results only if RTTs are small, i.e. when the stepping-stone chain is completely contained within one LAN-network.

DeepCorr, 2018 A more recent example of neural network application comes from Nasr et al. [20], who train a deep convolutional neural network to identify correlation between two connections from the upstream and downstream interarrival times and packet sizes in each connection. The trained network is large with over 200 input filters, and consists of three convolutional and three feed-forward layers. On stepping-stones, the authors achieve a 90% detection rate with 0.02% false positives.

RWalk, 2015 This model by Yang et al. [34] combines packet-counting methods and RTT mining methods to improve detection results from [32]. The authors improve widely-used packet-counting methods to resist chaff perturbation by counting the number of round-trips in a connection to determine if the connection is being relayed. Packet pairs representing a round-trip for each connection are estimated using a combination of packet matching and clustering, and counted as N_{in} and N_{out}.

C-Over, 2016 This method by Huang et al. [18] improves the detection methods proposed by Ding et al. [10]. Their detection model is built on the fact that in a long

connection chain, the round-trip-time of a packet may be longer than the intervals between two consecutive keystrokes. In a long connection chain, this will result in cross-overs between request and response, which causes the curve of sorted upstream RTTs to rise more steeply than in a regular connection.

Ano1, 2011 Crescenzo et al. [9] have proposed an assembly of three anomaly-based methods to detect time delays and chaff perturbations in a selected connection. Packet time-delays are detected by flagging acknowledgement packets if their response-time exceeds the average RTT. A distance-based chaff detection method compares the similarity of downstream with upstream packet sequences. Finally, a causality-based method verifies that the number of ON-intervals in upstream and downstream direction match. The authors claim successful detection for chaff ratios 25% or more, and for delays introduced to up to 70% of all packets.

Ano2, 2011/2013 Huang et al. [17] proposed an anomaly-based method to detect chaff perturbations since interarrival times in regular connections tend to follow a Pareto or Lognormal distribution, which chaffed connections supposedly do not. The authors perform a statistical test whether packets in a connection follow these distributions.A similar approach from the same authors [11] is directed towards the detection of packet jittering. The authors state 95% detection rate on connections with 50% chaff ratio and more while retaining zero false positives using a small set of chaffed interactive SSH stepping-stone connections.

WM, 2010 Watermarking typically yields very low false-positives for connection correlation. Wang et al. [28] provide an approach that offers at least some resistance against timing perturbations. The authors assume some limits to an adversary's timing perturbations, such as a bound on the delays, divide a sequence of packets into pairs, whose interarrival time is perturbed using a new watermarking function. The authors state 100% TP with 0.5% FP with resistance against timing perturbations of up to $1.4s$.

4.1 Implementation of selected approaches

In order to evaluate the above selected method on our generated data, we had to reimplement the algorithms according to the steps and design described in the corresponding publications. We implemented all methods in Python, and PyTorch if necessary.

For all methods, we either varied the classification threshold for AUC calculation, or estimated it to yield a specific false positive rate across all methods. For anomaly-based methods, we had to estimate multiple thresholds, for which we found the optimal combination via grid-search. For the WM approach, we set the quantization step size $s = 0.4$s, and the redundancy number $m = 11$ to consistently embed the watermark in the first 1500 packets. These values were also used by the authors.

Both DeepCorr and WuNeur are machine-learning based algorithms that require training. Since DeepCorr is a large model and requires a lot of training data, we train it using 80% of the evaluation data. WuNeur is sensitive to differing chain lengths, which is why we use 1000 connection pairs from each dataset BA, DA, CA, and CL, as well as 6000 connection pairs from the training data.

5 Results

5.1 Data without evasion tactics

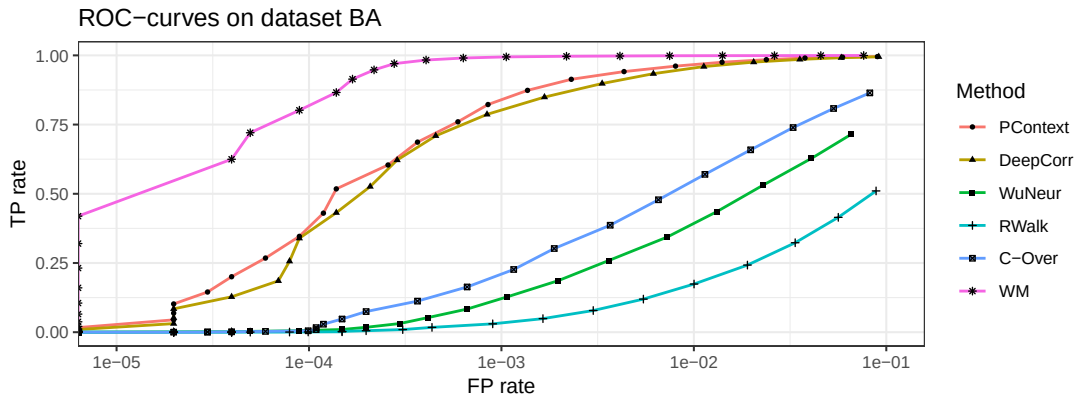

Fig. 4. ROC-curves for different SSD methods on dataset BA (no evasive tactics). Anomaly-based methods are excluded.

	PContext	DeepCorr	WuNeur	RWalk	C-Over	WM
AUC	0.998	0.997	0.938	0.853	0.965	0.9998

Table 3. AUC-scores for different methods on stepping-stone data without evasive tactics.

First, we look at the detection rates for traffic from stepping-stones that did not use any evasive tactics, i.e. $S_1, \ldots, S_4$ are only forwarding commands and responses. The successful detection of this activity with low false-positives should be the minimum requirement for any SSD method. Since anomaly-based approaches aim to only detect evasive behaviour, we exclude them from this analysis.

Fig. 4 depicts the calculated ROC-curves, which plot the true positive rate against the false positive rate for varying detection thresholds. Table 3 depicts the overall AUC-scores.

Unsurprisingly, the watermarking method achieves high detection results with very low false-positives. Both the PContext and DeepCorr models start to yield good detection results of around 80% at a FP rate lower than 0.1%, with the PContext method

slightly outpacing the DeepCorr method. RTT-based methods seem to not perform as well compared to the other included methods. Overall, the observed ROC curves seem to be in agreement with the stated detection rates of the selected methods except for RWalk.

5.2 Delays

We now consider the effect of transfer delays added by the attacker to packets on the detection rates. For that, we pick detection thresholds for each SSD methods corresponding to a FP rate of 0.4% as most methods are able to achieve at least moderate detection results at this rate. We look at delays added to only to outgoing packets on S_4, the last stepping stone in the chain. Fig. 5 depicts evolution of detection rates in dependence of the maximum delay δ_D.

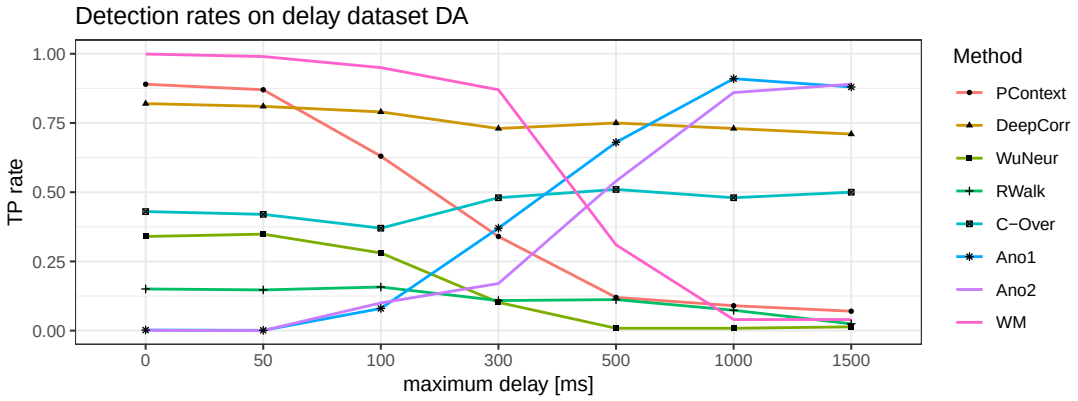

Fig. 5. Detection rates in dependence of δ_D for different methods on dataset DA with a fixed FP rate of 0.4%.

As visible, both anomaly-based methods are capable of detecting added delays relatively reliably above a certain threshold. Furthermore, both the detection rates of DeepCorr and the RTT-based C-Over only decrease slightly under the influence of delays. Detection rates for all other methods decrease significantly to the point where no meaningful predictions can be made. This is also reflected by the AUC-scores for traffic with $\delta_D = 1000ms$, given in Table 4.

While the WM method is robust against transfer delays up to $\delta_D = 500$ms, this value is smaller than the one claimed by the authors. This might however be a result of the slightly smaller quantisation step size that we used. It is surprising that the PContext method shows only little robustness against transfer delays, which contradicts the authors claims, potentially due to the incorrect assumption that relying on *Echo*-packets are not subject to transfer delays.

	PContext	DeepCorr	WuNeur	RWalk	C-Over	Ano1	Ano2	WM
AUC	0.638	0.995	0.613	0.641	0.952	0.997	0.996	0.562

Table 4. AUC-scores for SSD methods with added transfer delays at $\delta_D = 1000ms$.

5.3 Chaff

We now consider the effect of chaff perturbations added by the attacker to individual connections on the detection rates. Again we pick detection thresholds for each SSD methods corresponding to a FP rate of 0.4%.

Chaff packets are added to both the connection between S_3 and S_4 as well as between S_4 and host T as described in Section 2.3. Fig. 6 depicts evolution of detection rates in dependence of the ratio of number of chaff packets to packets from the actual interaction.

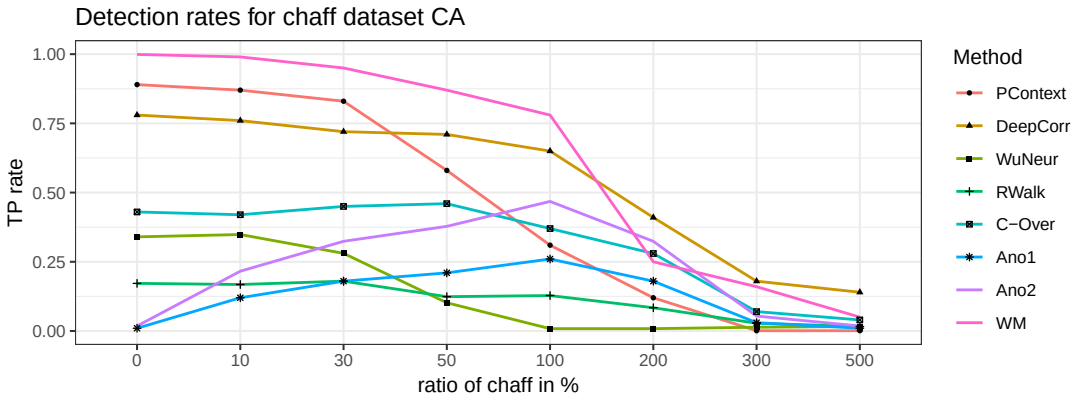

Fig. 6. Detection rates in dependence of δ_C for different methods on dataset CA with a fixed FP rate of 0.4%

As visible, all methods struggle to detect stepping stones once the chaff packets become the majority of the transferred traffic. This is also evident from the AUC-scores given in Table 4. Several approaches claimed to be resistent to chaff perturbations, however prior evaluations were limited chaff ratios below 100% without obvious reason.

It is surprising that the anomaly detection methods do not perform better at detecting chaff perturbations. Chaff in both approaches was however evaluated with different traffic generation distribution and not compared against a background of traffic following a similar generation distribution, which could explain the disagreement between the results we are finding here.

Overall, these results are in disagreement with the "robustness" claims made for four of the selected approaches, namely PContext, RWalk, Ano1, and Ano2.

	PContext	DeepCorr	WuNeur	RWalk	C-Over	Ano1	Ano2	WM
AUC	0.639	0.886	0.615	0.641	0.589	0.782	0.738	0.839

Table 5. AUC-scores for SSD methods with added chaff at 300% ratio.

5.4 Summary

Overall, detection rates on dataset BA are mostly in line with the claimed capabilities except for RWalk, although detection rates are slightly lower than stated by most authors. Delay perturbation increases detection difficulty for most methods, except for Ano1, Ano2, and DeepCorr, which contradicts robustness claims for PContext and to some extend WM. Our inserted chaff perturbations however render detection impossible for all methods examined, which contradicts robustness claims for PContext, Ano1, Ano2, and RWalk, even though the claims were based on lower chaff levels.

As discussed in Section A and B, longer chains yield higher detection rates for RTT-based methods while Different network transmission settings seem to have overall little influence on detection rates.

6 Related work

6.1 Testbeds and data

In 2006, Xin et al. [31] developed a standard test bed for stepping-stone detection, called *SST*, with the goal to enable a reproducible evaluation of stepping stone chain detection algorithms with easy configuration and operation. The tool allows for an arbitrary number of intermediate hosts and generates scripts to mimic interactive SSH and TelNet connections. In contrast to our work, the authors give little detail on implemented evasive tactics. To our knowledge, SST has only been used for evaluation by Zhang et al. [35], and is not available anymore.

An approach to use publicly available data comes from Houmansadr et al. [20]. The authors simulate stepping stone behaviour using the well-known CAIDA anonymised data traces [4] by adding packet delays and drops retroactively on selected connections. While this procedure seems sufficient for the evaluation of watermarking methods, it falls short on simulating the effects of an actual connection chain and leaves out chaff perturbations.

We find that when authors evaluate methods on self-generated data, tested evasive behaviours are often lacking analytical discussion and their implementations are too simplistic, leading to increased detection rates. An example of this can be seen in the evaluation of Ano1 [9], where a standard option in netcat is used to generate chaff perturbations for evaluation, or for PContext [34] where simulated chaff is added randomly after the traffic collection.Furthermore, often a relatively low limit on the amount of inserted chaff perturbations is assumed without obvious reason, thus avoiding evaluation at higher ratios.

6.2 Surveys

Wang et al. [27] recently conducted an extensive survey of stepping stone intrusion detection. The authors group methods according to the respective methodology but do not cover graph-based methods such as [15], or anomaly-based methods such as [9].Shullich et al. [23] in 2011 also conducted a survey on stepping stone intrusion detection. The authors perform a similar grouping of methods, but also discuss related work in evasion tactics and test frameworks. The authors furthermore give an outlook on areas for future research, such as hacker motivation, the cardinality problem, or the difficulty of tracing back chains through firewalls.Neither survey provide an evaluation of detection rates nor a direct comparison of the rates achieved by the corresponding authors.

6.3 Areas not covered in this work

Graph-based and behavioural models do not examine individual connections but instead focus on the overall temporal activity of one or more hosts in a network. Notable examples include Apruzzese et al. [5] and Gamarra et al. [15], who both developed graph-based models . Simulating temporal stepping-stone activity realistically is difficult, with little empirical data to support specific assumptions. Since our dataset focuses only on individual connections and corresponding evasion, we are not able to include graph-based or behavioural approaches in our evaluation.

Coskun et al. [8] identify another form of stepping-stones called *store-and-forward*, which transfer data within files in a non-interactive manner. Though harder to detect than interactive connections, this procedure limits the attackers ability to explore the target, which is why SSD research has been primarily concerned with interactive stepping-stones.

7 Conclusion

In this work, we set out to evaluate the state-of-the-art of SSD methods using a comprehensive data generation framework. Our framework simulates realistic stepping-stone behaviour with SSH-tunnels in different settings. The implemented evasive perturbation tactics follow the findings of Padhye et al. [21] to resemble streaming services, and are adjustable in volume. We will release a large dataset that highlights multiple aspects in SSD, and is suitable to train ML-based methods.

Overall, our results show that attackers can reliably evade detection by using the right type and amount of chaff perturbation, which disproves several claims made about the robustness against this evasive tactic. Although to a lesser degree, our implemented delay perturbations still affect detection rates for most methods, but can be detected by anomaly-based methods.

Currently, it seems that watermarking methods are currently most suited to reliably detect simple stepping-stones in real-life deployment. The performance of DeepCorr indicates that deep neural networks show the most potential at detecting attacks that

use chaff or delay perturbations if they are trained on suitable data. We find that detection and false-positive rates for RTT-based methods are significantly lower than for other methods, which makes them less suitable as stand-alone solutions.

References

1. M-trends 2015: A view from the front lines. Technical report, 2015.
2. Mcafee technical report on night dragon operation. Technical report, 2015.
3. Archimedes documentation, 2017.
4. The CAIDA UCSD Anonymized Internet Traces 2018, 2018. Accessed: 2020-02-10.
5. Giovanni Apruzzese, Fabio Pierazzi, Michele Colajanni, and Mirco Marchetti. Detection and threat prioritization of pivoting attacks in large networks. *IEEE Transactions on Emerging Topics in Computing*, 2017.
6. Luis Ayala. Active medical device cyber-attacks. In *Cybersecurity for Hospitals and Healthcare Facilities*, pages 19–37. Springer, 2016.
7. Henry Clausen, Robert Flood, and David Aspinall. Traffic generation using containerization for machine learning. In *Proceedings of the Dynamic and Novel Advances in Machine Learning and Intelligent Cyber Security Workshop*. ACM, 2019.
8. Baris Coskun and Nasir Memon. Efficient detection of delay-constrained relay nodes. In *Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007)*, pages 353–362. IEEE, 2007.
9. Giovanni Di Crescenzo, Abhrajit Ghosh, Abhinay Kampasi, Rajesh Talpade, and Yin Zhang. Detecting anomalies in active insider stepping stone attacks. *JoWUA*, 2(1):103–120, 2011.
10. Wei Ding, Matthew J Hausknecht, Shou-Hsuan Stephen Huang, and Zach Riggle. Detecting stepping-stone intruders with long connection chains. In *2009 Fifth International Conference on Information Assurance and Security*, volume 2, pages 665–669. IEEE, 2009.
11. Wei Ding, Khoa Le, and Shou-Hsuan Stephen Huang. Detecting stepping-stones under the influence of packet jittering. In *2013 9th International Conference on Information Assurance and Security (IAS)*, pages 31–36. IEEE, 2013.
12. David L Donoho, Ana Georgina Flesia, Umesh Shankar, Vern Paxson, Jason Coit, and Stuart Staniford. Multiscale stepping-stone detection: Detecting pairs of jittered interactive streams by exploiting maximum tolerable delay. In *International Workshop on Recent Advances in Intrusion Detection*, pages 17–35. Springer, 2002.
13. EU ENISA. Baseline security recommendations for iot in the context of critical information infrastructures, 2017.
14. Gordon Fraser. Tunneling, pivoting, and webapplication penetrationtesting. Technical report, SANS, 2015.
15. Marco Gamarra, Sachin Shetty, David M Nicol, Oscar Gonazlez, Charles A Kamhoua, and Laurent Njilla. Analysis of stepping stone attacks in dynamic vulnerability graphs. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2018.
16. Stephen Hemminger et al. Network emulation with netem. In *Linux conf au*, pages 18–23, 2005.
17. Shou-Hsuan Stephen Huang and Ying-Wei Kuo. Detecting chaff perturbation on stepping-stone connection. In *2011 IEEE 17th International Conference on Parallel and Distributed Systems*, pages 660–667. IEEE, 2011.
18. Shou-Hsuan Stephen Huang, Hongyang Zhang, and Michael Phay. Detecting stepping-stone intruders by identifying crossover packets in ssh connections. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 1043–1050. IEEE, 2016.
19. Robert M. Lee, Michael J. Assante, and Tim Conway. Analysis of the cyber attack onthe ukrainian power grid. Technical report, E-ISAC, 2016.
20. Milad Nasr, Alireza Bahramali, and Amir Houmansadr. Deepcorr: Strong flow correlation attacks on tor using deep learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1962–1976, 2018.

21. Jaideep D Padhye, Kush Kothari, Madhu Venkateshaiah, and Matthew Wright. Evading stepping-stone detection under the cloak of streaming media with sneak. *Computer Networks*, 54(13):2310–2325, 2010.
22. Vern Paxson and Sally Floyd. Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on networking*, 3(3):226–244, 1995.
23. Robert Shullich, Jie Chu, Ping Ji, and Weifeng Chen. A survey of research in stepping-stone detection. *" International Journal of Electronic Commerce Studies"*, 2(2):103–126, 2011.
24. Stuart Staniford-Chen and L Todd Heberlein. Holding intruders accountable on the internet. In *Proceedings 1995 IEEE Symposium on Security and Privacy*, pages 39–49. IEEE, 1995.
25. Colin Tankard. Advanced persistent threats and how to monitor and deter them. *Network security*, 2011(8):16–19, 2011.
26. Petr Velan, Jana Medková, Tomáš Jirsík, and Pavel Čeleda. Network traffic characterisation using flow-based statistics. In *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*, pages 907–912. IEEE, 2016.
27. Lixin Wang and Jianhua Yang. A research survey in stepping-stone intrusion detection. *EURASIP Journal on Wireless Communications and Networking*, 2018(1):276, 2018.
28. Xinyuan Wang and Douglas Reeves. Robust correlation of encrypted attack traffic through stepping stones by flow watermarking. *IEEE Transactions on Dependable and Secure Computing*, 8(3):434–449, 2010.
29. Han-Ching Wu and Shou-Hsuan Stephen Huang. Performance of neural networks in stepping-stone intrusion detection. In *2008 IEEE International Conference on Networking, Sensing and Control*, pages 608–613. IEEE, 2008.
30. Han-Ching Wu and Shou-Hsuan Stephen Huang. Neural networks-based detection of stepping-stone intrusion. *expert systems with applications*, 37(2):1431–1437, 2010.
31. Jianqiang Xin, Lingeng Zhang, Brad Aswegan, John Dickerson, T Daniels, and Yong Guan. A testbed for evaluation and analysis of stepping stone attack attribution techniques. In *2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, 2006. TRIDENTCOM 2006.*, pages 9–pp. IEEE, 2006.
32. Jianhua Yang and Shou-Hsuan Stephen Huang. Mining tcp/ip packets to detect stepping-stone intrusion. *computers & security*, 26(7-8):479–484, 2007.
33. Jianhua Yang and David Woolbright. Correlating tcp/ip packet contexts to detect stepping-stone intrusion. *Computers & Security*, 30(6-7):538–546, 2011.
34. Jianhua Yang and Yongzhong Zhang. Rtt-based random walk approach to detect stepping-stone intrusion. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, pages 558–563. IEEE, 2015.
35. Linfeng Zhang, Anthony Persaud, Alan Johnson, and Yong Guan. Stepping stone attack attribution in non-cooperative ip networks. In *Proc. of the 25th IEEE International Performance Computing and Communications Conference (IPCCC 2006)*, 2005.

.1 False positives

Table 6 depicts the relative contribution[2] at $FP = 0.4\%$ of each of the three benign data types to the overall false positive rate. Most methods have more problems with the heterogeneous nature the CAIDA traces, with only PContext and DeepCorr seeing most false positives in the SSH traffic.

The multimedia traffic is causing most problems for the anomaly-based methods, persumably because it follows a similar distribution as the generated chaff perturbations.

[2] after adjusting for their weight

	PContext	DeepCorr	WuNeur	RWalk	C-Over	Ano1	Ano2	WM
CAIDA	0.36	0.46	0.47	0.67	0.53	0.48	0.35	0.81
SSH	0.53	0.46	0.21	0.28	0.27	0.05	0.02	0.08
multimedia	0.11	0.08	0.32	0.04	0.20	0.47	0.63	0.11

Table 6. Relative contribution in % of different benign data to the FP rate.

A Influence of chain length

In this section, we look at the effect of differing chain lengths on the detection rates. We only focus on RTT-based methods here since the other methods should and do not see a significant effect from varying chain lengths[3]. Since RTT-based methods aim to measure the effect of packets travelling via multiple hosts, it is unsurprising that they perform better at detecting longer chains.

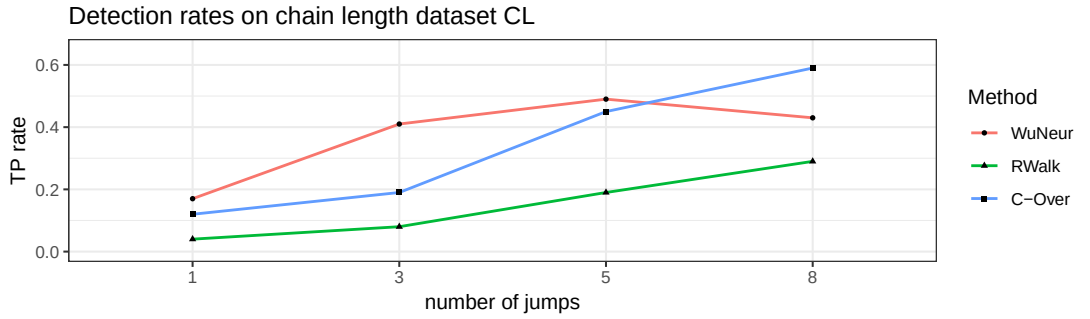

Fig. 7. Detection rates in dependence of chain length for different methods on dataset CL with a fixed FP rate of 0.4%

Of the RTT-based methods, only C-Over was able to yield consistent detection rates under transfer delays. Interestingly, if the C-Over method is applied to connections between S_3 and S_4 instead of between S_4 and the target, detection rates decrease in the same manner as for other RTT-based methods. This is not surprising as the underlying assumption for robustness for this approach relies on Echo-packets not being delayed.

B Influence of network settings

Finally, we look at the effect of different nework settings. We only show methods that show significant effects and omitted bandwidth from the evaluation as different values do not seem to have any effect on detection rates[4].

[3] For non-RTT-methods, the detection rate error ($2.6\% - 6.5\%$) for each length was larger than the detection rate differences ($0.2\% - 3.7\%$) across different lengths.

[4] For all methods, the detection rate differences ($0.7\% - 6.2\%$) were smaller across bandwidths than the overall detection rate errors ($2.6\% - 6.5\%$).

	Value	TP deviation from average				
		DeepCorr	WuNeur	RWalk	C-Over	WM
RTT	5ms	−0.2%	+41.3%	−42.3%	−36%	+0.03%
	70ms	−5.6%	−5.8%	+35.1%	+51%	−2.2%
Packet loss	0%	+1.2%	+1.3%	+2.1%	+4.3%	+0.02%
	7%	−9.1%	−1.1%	−3.1%	−7.3%	−9.7%

Table 7. Influence of network congestion on detection rates at a fixed FP rate of 0.4%. The given percentages are describing the change of the detection rate under the given congestion setting when compared to the overall average.

As visible in Table 7, the three RTT-based methods show different responses to small/large average round-trip-times. While WuNeur, as expected from prior results, performs better in LAN settings, detection rates of the RWalk and C-Over methods are boosted by larger RTTs. All methods profit from lower packet losses.