ROBUST DETECTION OF STEPPING-STONE ATTACKS

Ting He and Lang Tong∗
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853, USA
Email:{th255,lt35}@cornell.edu

***Abstract*— The detection of encrypted stepping-stone attack is considered. Besides encryption and padding, the attacker is capable of inserting chaff packets and perturbing packet timing and transmission order. Based on the assumption that packet arrivals form renewal processes, and a pair of such renewal processes is also renewal, a nonparametric detector is proposed to detect attacking traffic by testing the correlation between interarrival times in the incoming process and the outgoing process. The detector requires no knowledge of the interarrival distributions, and it is shown to have exponentially decaying detection error probabilities for all distributions. The error exponents are characterized using the Vapnik-Chervonenkis Theory. An efficient algorithm is proposed based on the detector structure to detect renewal processes with linearly correlated interarrival times. It is shown that the proposed algorithm is robust against an amount of chaff arbitrarily close to the amount of chaff needed to mimic independent processes.**

***Keywords:*Intrusion detection, Stepping-stone attacks, Statistical Learning Theory, Nonparametric detection.**

I. Introduction

Stepping-stone attack is a common way of launching anonymous attacks [1]. In such an attack, the attacker routes attacking packets to the victim through a chain of compromised hosts called "stepping stones". The victim only sees the last stepping stone, and thus the attacker's identity is concealed. The difficulty in defending against such attacks lies in the tracing of the attacking path, and the tracing can be decomposed into detecting every pair of stepping-stone connections on the intrusion path.

A sophisticated attacker can modify the attacking traffic to thwart detection. In particular, he can encrypt and pad the packets so that no information is revealed by the bit patterns or the lengths of packets; the only information available to the detector is the timing of the traffic. The timing, however, is subject to changes introduced by the attacker such as random delay and packet reshuffling. Furthermore, the attacker can mix attacking traffic with chaff—dummy traffic generated purely for the purpose of evading detection. Chaff traffic can be generated arbitrarily, and it does not need to reach the victim.

In this paper, we consider the problem of detecting encrypted stepping-stone connections in the presence of chaff. We allow the attacker to use various evasion strategies including encryption, padding, changing the packet order and timing, and mixing attacking packets with chaff. Our goal is to develop techniques that are robust against the presence of chaff, forcing the attacker to spend a substantial amount of time transmitting chaff. Such robust techniques coupled with constrains on rates may be one way to minimize the effectiveness of the attacker.

A. Related Work

Ever since Staniford and Heberlein [1] first consider the problem of detecting stepping-stone connections, there has been a continuous evolution of detection techniques as well as evasion strategies. Early content-based detection techniques such as [1], [2] are easily defeated by encryption and padding. Timing-based detection considered in [3]–[5] is not affected by encryption or padding, but is vulnerable to active timing perturbation introduced by the attacker.

Donoho *et al.* [6] first consider the randomly delayed stepping-stone connections, and since then a number of advances have been made in detecting encrypted, transformed stepping-stone connections; see [6]–[8]. The key assumption of these methods is that there is a limit on the attacker's ability to alter the traffic. For example, Donoho *et al.* [6] show that in principle it is possible to detect transformed Poisson processes if the transformation satisfies a bounded delay. Wang and Reeves in [7] propose to correlate relayed streams with independent and identically distributed, order-preserving perturbation by introducing watermarks into packet interarrival times. Blum *et al.* [8] present an algorithm "DETECT-ATTACKS" (DA), the first passive detection algorithm with guaranteed performance based on the assumption of bounded delay and bounded peak rate.

When chaff can be inserted to evade detection, many previous algorithms fail. Blum *et al.* [8] propose an algorithm called "DETECT-ATTACKS-CHAFF" (DAC) modified from their algorithm DA to deal with limited chaff. Algorithm DAC tolerates a fixed number of chaff packets by sacrificing the false alarm probability, but a pair of arbitrarily long streams can still evade detection by adding a constant number of chaff packets. Peng *et al.* [9] and Zhang *et al.* [10] separately

This work is supported in part by the National Science Foundation under award CCF-0635070, the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program DAAD19-01-2-0011, and TRUST (The Team for Research in Ubiquitous Secure Technology) sponsored by the National Science Fundation under award CCF-0424422. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

propose packet-matching schemes for robust detection; it, however, turns out that these schemes can not deal with chaff packets in the incoming process at all.

All of the schemes in [8]–[10] can be defeated by a constant number of chaff packets. As the traffic size increases, the fraction of chaff will go to zero. In terms of rate, zero rate chaff traffic suffices to evade their detection. The only algorithms that are known to handle chaff traffic of non-zero rate are algorithms “DETECT-BOUNDED-DELAY-CHAFF” (DBDC) and “DETECT-BOUNDED-MEMORY-CHAFF” (DBMC) in [11]. Algorithm DBDC can detect traffic flows with up to $1/(1+\lambda\Delta)$ fraction of chaff (where λ is a design parameter) if packet delays are bounded by Δ. Algorithm DBMC is designed for detecting traffic flows through a host which can hold at most M packets, and is robust against up to $1/(1+M)$ fraction of chaff. The drawback of DBDC and DBMC is that the false alarm probabilities, although shown to go to zero eventually, can be large for finite sample size. In this paper, we want to answer the question whether it is possible to reduce the false alarm probability by allowing certain miss detection.

B. Summary of Results and Organization

We consider robust detection of stepping-stone connections in the presence of chaff. To the best of our knowledge, no existing detector has provable decay rate in the probabilities of both false alarm and miss detection. The main contribution of this paper is a quantitive characterization of both false alarm and miss probabilities by imposing the assumption that pairs of interarrival times in the incoming and outgoing processes are independent and identically distributed (*i.i.d.*). The *i.i.d.* assumption is a limiting assumption in the sense that even if *i.i.d.* perturbation is applied to a renewal process, the generated pair of processes may not have *i.i.d.* interarrivals; it is, however, general enough to include a wide range of relayed processes because we do not assume the processes to satisfy any other statistical property. The stepping-stone detector should therefore be nonparametric.

We propose a nonparametric detector to detect renewal processes with correlated interarrival times based on the assumption that the pair formed by these renewal processes is also renewal, *i.e.*, the pairs of interarrival times from the incoming and outgoing processes are *i.i.d.* ; the detector does not assume the knowledge of the interarrival distributions. This detector applies to general attacking traffic with or without memory or delay constraints. We show that the probabilities of miss detection and false alarm both decay exponentially with the number of packets used in the detection. Explicit expressions of the error exponents are given using the Vapnik-Chervonenkis (VC) Theory. Such expressions allow us to design the detector threshold to satisfy prescribed performance specifications. The proposed detector is optimal under the renewal assumption in the sense that if the attacking packets satisfy the bounded memory or bounded delay constraint, then the amount of chaff needed to evade detection is proportional to the traffic size, and the proportion can be arbitrarily close to what is needed to mimic truly independent processes. An algorithm is proposed to efficiently implement the detector; it reduces the computation complexity from $O(n^6)$ to $O(n^2 \log n)$ where n is the sample size.

The rest of the paper is organized as follows. Section II defines the problem. Section III presents a nonparametric detector to deal with chaff, and analyzes its performance. The section also presents an efficient algorithm to implement the detector. Section IV compares the robustness of the proposed detector with that of existing stepping-stone detectors. Section V simulates the proposed detector for pairs of renewal processes with bivariate exponential interarrival distributions. Then Section VI concludes the paper with comments on some practical issues about the application of such a detector.

II. Problem Definition

Denote the packet arrivals on stream i as a point process

$$S_i = (s_1^{(i)},\, s_2^{(i)},\, s_3^{(i)}, \ldots), \quad i = 1,\, 2, \ldots$$

where $s_k^{(i)}$ $(k \geq 1)$ is the kth arrival epoch in stream i. Let $\mathcal{T}_i = \{s_1^{(i)},\, s_2^{(i)}, \ldots\}$ be the set of the elements in S_i. Let $(S_1,\, S_2)$ be a pair of incoming and outgoing streams of interest at a particular gateway node. Normally, S_1 and S_2 are independent. If, however, S_2 is a relayed stream of S_1, then they will satisfy certain relations.

Definition 2.1: A pair of streams $(S_1,\, S_2)$ is a *normal pair* if S_1 and S_2 are independent point processes. It is a *stepping-stone pair* if there exists a bijection $g : \mathcal{T}_1 \rightarrow \mathcal{T}_2$ such that $g(s) - s \geq 0$ for any $s \in \mathcal{T}_1$, and g satisfies certain communication requirements.

The bijection g, unknown to the detector, is a mapping between the arrival and the departure epochs of the same packets, allowing permutation of packets during the relay. The condition that g is a bijection imposes a *packet conservation* constraint, *i.e.*, no attacking packets are generated or dropped at the stepping stones. The condition $g(s)-s \geq 0$ is the *causality* constraint, which means that an attacking packet cannot leave a host before it arrives. Communication requirements are due to the need of the attacker’s application, the physical constraints of the relay host, or the communication channel. Examples include, but are not limited to, bounded memory constraint and bounded delay constraint; see [11].

If S_i $(i = 1, 2)$ is the mixture of attacking packets and chaff, then the requirements are relaxed, as stated in the following definition.

Definition 2.2: A pair of streams $(S_1,\, S_2)$ is a *stepping-stone pair with chaff* if it is the superposition of a stepping-stone pair $(S_1',\, S_2')$ and a pair of arbitrary streams $(C_1,\, C_2)$[1].

Stream C_i $(i = 1, 2)$ consists of dummy packets called *chaff* which do not need to arrive at the destination. Chaff packets can be generated or dropped at any stepping stones without affecting the attack.

Let the interarrival times of S_1 be $X_1,\ X_2, \ldots$, where $X_1 = s_1^{(1)}$, and $X_i = s_i^{(1)} - s_{i-1}^{(1)}$ $(i > 1)$. Similarly, denote the interarrival times of S_2 by $Y_1,\ Y_2, \ldots$. If all the transmissions in the network follow renewal processes, then X_i's and Y_i's are *i.i.d.* , respectively. The problem is that without any constraint on stepping-stone pairs, $(X_i)_{i=1,\,2,\ldots}$ and $(Y_i)_{i=1,\,2,\ldots}$ may correlate arbitrarily; in general, samples of the pairs $(X_i,\ Y_i)$ $(i = 1,\ 2, \ldots)$ are not sufficient for detection because the order in which these samples are taken are also relevant. The hypothesis testing will have the form of

$$\begin{aligned} \mathcal{H}_0: &\qquad P(\mathbf{X}^n,\, \mathbf{Y}^n) = P(\mathbf{X}^n)P(\mathbf{Y}^n), \\ \mathcal{H}_1: &\qquad P(\mathbf{X}^n,\, \mathbf{Y}^n) \neq P(\mathbf{X}^n)P(\mathbf{Y}^n), \end{aligned}$$

for any $\mathbf{X}^n,\ \mathbf{Y}^n \in \mathbb{R}^{+n}$. For arbitrary stepping-stone pairs, the worst case complexity grows exponentially with the sample size. If, however, the stepping-stone pairs are renewal as well, *i.e.,* the pairs $(X_i,\ Y_i)$ $(i = 1,\ 2, \ldots)$ are *i.i.d.* , then the detection is reduced to a testing of the following single-lettered hypotheses[2]:

$$\mathcal{H}_0:\ P_{XY} = P_X \circ P_Y, \qquad \mathcal{H}_1:\ P_{XY} \neq P_X \circ P_Y, \qquad (1)$$

given realizations of $((X_1,\ Y_1),\ (X_2,\ Y_2), \ldots)$. This is a nonparametric hypothesis testing problem; no specific assumptions on the distribution P_{XY} are imposed.

III. Nonparametric Detection of Renewal Traffic

Donoho *et al.* in [6] have noticed that for renewal processes, local timing perturbation or reshuffling will not destroy the correlation between processes. Furthermore, they show that nonzero correlation can be obtained even if the attacker inserts chaff independent of the attacking traffic. Although Donoho *et al.* do not derive specific stepping-stone detectors in [6], their work shows that, in principle, effective detection can be achieved in the presence of chaff. Inspired by Donoho *et al.* [6], we propose an alternative to existing algorithmic approaches that check strict memory or delay constraints. We aim at deriving a detector to test the statistical correlation between processes. It is desirable that the detector has guaranteed performance for a wide range of traffic.

In this section, we present a nonparametric detector based on the statistical learning theory for the hypothesis testing problem defined in (1). In Section III-A, we introduce a distance measure, called $\mathcal{A}$-distance, between probability distributions, and define a detector based on $\mathcal{A}$-distance. We then address the computation issues in Section III-B, where an efficient algorithm is proposed to reduce the complexity in implementing the $\mathcal{A}$-distance detector.

A. Distance Measure and Detector

To test $\mathcal{H}_0$ against $\mathcal{H}_1$, we need to measure the distance between probability distributions. In a parametric framework, the conventional distance measure is the Kullback-Leibler distance [12]. Under the nonparametric framework, however, the Kullback-Leibler distance cannot be easily replaced by its finite sample counterpart[3] We solve this problem by using the following pseudo distance measure from [13]:

Definition 3.1 ($\mathcal{A}$-distance and empirical $\mathcal{A}$-distance): Given probability spaces[4] $(X, \mathcal{F}, P_i)$ $(i = 1,\ 2)$ and a collection of sets $\mathcal{A} \subseteq \mathcal{F}$, the *$\mathcal{A}$-distance* between P_1 and P_2 is defined as

$$d_{\mathcal{A}}(P_1, P_2) = \sup_{A \in \mathcal{A}} |P_1(A) - P_2(A)|.$$

Given two collections of samples S_1, S_2 drawn independently and i.i.d. from P_1, P_2 respectively, the *empirical $\mathcal{A}$-distance* $d_{\mathcal{A}}(S_1, S_2)$ is similarly defined by replacing $P_i(A)$ with the empirical probability

$$S_i(A) \stackrel{\Delta}{=} \frac{|S_i \bigcap A|}{|S_i|},$$

where $|S_i \cap A|$ is the number of samples from S_i that are in the set A.

We see that $d_{\mathcal{A}}(S_1,\ S_2) \in [0,\ 1]$. By Vapnik-Chervonenkis Inequality [14], it is shown [15] that $d_{\mathcal{A}}(S_1,\ S_2)$ can be arbitrarily close to $d_{\mathcal{A}}(P_1,\ P_2)$ as sample size goes to infinity.

Given samples $S = \{(x_i,\ y_i)\}_{i=1}^n$, let $S_X = \{x_i\}_{i=1}^n$, and $S_Y = \{y_i\}_{i=1}^n$. With the distance measure defined, we now specify the detector as follows:

Definition 3.2: Let $\mathcal{A}$ be a collection of measurable subsets of $[0,\ \infty) \times [0,\ \infty)$. Given $\epsilon \in (0,\ 1)$, the detector using $\mathcal{A}$-distance measure to test the hypotheses in (1) is defined as[5]

$$\delta_{d_{\mathcal{A}}}(S,\ \epsilon) = \begin{cases} 1 & \text{if } d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY}) > \epsilon, \\ 0 & \text{o.w.,} \end{cases}$$

[1] Note that C_1 and C_2 may not have equal length, and either of them can be empty.

[2] We use $P_X \circ P_Y$ to denote the joint probability distribution for $(X,\ Y)$ in which X and Y are independent with marginals P_X and P_Y, respectively.

[3] For example, it can be shown that for continuous distribution, the empirical Kullback-Leibler distance is infinite almost surely.

[4] We use the convention that X is the sample space, $\mathcal{F}$ the σ-field, and P_i the probability measure.

[5] We use the convention that the detector gives the value 1 for $\mathcal{H}_1$, and 0 for $\mathcal{H}_0$.

where $d_{\mathcal{A}}(S_X \circ S_Y, S_{XY})$ is the empirical $\mathcal{A}$-distance between $S_X \circ S_Y$ and S_{XY}, defined as

$$d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY}) = \sup_{A \in \mathcal{A}} \left| S_X \circ S_Y(A) - S_{XY}(A) \right|,$$

with $S_X \circ S_Y(A) \overset{\Delta}{=} |(S_X \times S_Y) \cap A|/|S|^2$, and $S_{XY}(A) \overset{\Delta}{=} |S \cap A|/|S|$.

The definition involves calculating the supremum over a possibly infinite collection of sets. The computation of the statistics will be addressed in Section III-B.

B. Efficient Computation of Test Statistics

Here we address the issue of computing the test statistics $d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY})$ defined in Definition 3.2. We give an algorithm to compute $d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY})$ efficiently for the class of bands tilted to a certain angle.

Consider $\mathcal{A}$ as the class of bands tilted to 45° with respect to the x-axis, *i.e.*, $A \in \mathcal{A}$ is of the form $\{(x,\ y) : y - x \in [a,\ b]\}$ for some $a \leq b$. The rationale for this choice of $\mathcal{A}$ is that in stepping-stone connections, the nth arrival $\sum_{i=1}^{n} X_i$ and the nth departure $\sum_{i=1}^{n} Y_i$ will not diverge unboundedly, so we expect $X_i \approx Y_i$. Thus the samples of interarrival pairs from stepping-stone traffic often cluster around the unit line $x = y$ with some noise; bands around the unit line can reveal significant difference between normal traffic and stepping-stone traffic.

Given a set of samples $S = \{(x_i,\ y_i)\}_{i=1}^{n}$, the "product samples" $S_X \times S_Y$ are the set of the n^2 points $\{(x_i, y_j)\}_{i,\ j=1}^{n}$. Sort the "product samples" into $(s'_1,\ s'_2, \ldots,\ s'_{n^2})$, where $s'_k = (x'_k, y'_k)$, such that $x'_1 - y'_1 \leq x'_2 - y'_2 \leq \ldots$. Geometrically, this sorting allows us to scan the "product samples" in the order they cross the 45° line as it moves from northwest to southeast, as illustrated in Fig. 1. Let $B(k,\ l)$ ($k \leq l$) be the 45° band with boundaries passing through s'_k and s'_l respectively (*e.g.*, $B(2,\ 6)$ in Fig. 1), *i.e.*,

$$B(k,\ l) \overset{\Delta}{=} \{(x,\ y) :\ x - y \in [x'_k - y'_k,\ x'_l - y'_l]\}.$$

We have that

$$\begin{aligned} & d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY}) \\ & \quad = \max_{1 \leq k \leq l \leq n^2} \left| \frac{|(S_X \times S_Y) \cap B(k,\ l)|}{|S|^2} - \frac{|S \cap B(k,\ l)|}{|S|} \right| \end{aligned}$$

For $|S| = n$, an exhaustive search to compute $d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY})$ will take $O(n^6)$ time, since there are $O(n^4)$ $B(k,\ l)$'s, and the computation for each $B(k,\ l)$ takes $O(n^2)$ time. By proper updating, however, we can reduce the complexity to $O(n^2 \log n)$ as shown in the algorithm below.

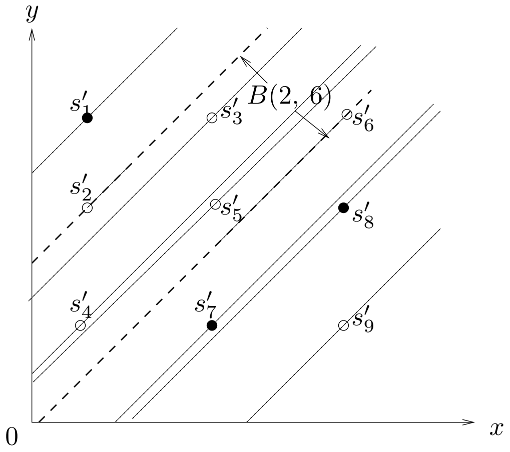

Fig. 1. Example: $n = 3$. ●: sample; ○: "product sample"; $B(2,\ 6)$: the 45° band between s'_2 and s'_6.

1) SEARCH-TILTED-BANDS (STB): Algorithm STB implements the $\mathcal{A}$-distance detector for the class of 45° bands efficiently. Define

$$F(k) \overset{\Delta}{=} \frac{|(S_X \times S_Y) \cap B(1,\ k)|}{|S|^2} - \frac{|S \cap B(1,\ k)|}{|S|},$$

for $k = 1, \ldots,\ n^2$, and $F(0) = 0$. Then we have that

$$\begin{aligned} d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY}) &= \max_{0 \leq k < l \leq n^2} |F(l) - F(k)| \\ &= \max_{0 \leq k \leq n^2} F(k) - \min_{0 \leq k \leq n^2} F(k). \quad (2) \end{aligned}$$

Algorithm STB computes $d_{\mathcal{A}}(S_X \circ S_Y,\ S_{XY})$ by computing $F(k)$ efficiently. The algorithm is shown in Table I.

TABLE I
SEARCH-TILTED-BANDS (STB)

```
SEARCH-TILTED-BANDS($S$, $\epsilon$):
  for $i,\ j = 1 : n$
    $D((j-1)n + i) = x_i - y_j$;
  end;
  $[\tilde{D}, I] = \text{sort}(D)$;
  $F_{\min} = F_{\max} = F(0) = 0$;
  for $k = 1 : n^2$
    $F(k) = \begin{cases} F(k-1) + \frac{1}{n^2} - \frac{1}{n} & \text{if } I(k) \bmod (n+1) == 1, \\ F(k-1) + \frac{1}{n^2} & \text{o.w.} \end{cases}$;
    $F_{\min} = \min(F_{\min},\ F(k))$;
    $F_{\max} = \max(F_{\max},\ F(k))$;
  end
  if $F_{\max} - F_{\min} > \epsilon$ return ATTACK;
  else return NORMAL;
```

In STB, I is an index array where $I(k)$ is the index of the kth smallest entry in D. If s'_k is the "product sample" corresponding to $D(I(k))$, we have that

$$F(k) = \begin{cases} F(k-1) + \frac{1}{n^2} - \frac{1}{n} & \text{if } s'_k \in S, \\ F(k-1) + \frac{1}{n^2} & \text{o.w.} \end{cases}$$

Note that $s'_k \in S$ if and only if $I(k) = (i-1)n + i = (i-1)(n+1) + 1$ for some $i \in \{1, \ldots,\ n\}$; therefore, $s'_k \in S$ is equivalent to $I(k)$ mod $(n+1) == 1$. Thus, STB can compute $F(k)$ ($k = 1, \ldots, n^2$) by an $O(n^2)$ updating. The sorting of D is the most time-consuming step, and it takes $O(n^2 \log n)$. Therefore, STB implements the $\mathcal{A}$-distance detector for the class of 45° bands in $O(n^2 \log n)$ time.

We point out that STB can be easily modified to detect other forms of linear correlation by changing the order in scanning the "product samples".

IV. Performance of $\mathcal{A}$-distance Detector

We now analyze the performance of $\delta_{d_\mathcal{A}}$. We show that it has exponentially decaying error probabilities on both false alarm and miss detection. We derive uniform upper bounds on the error probabilities by applying the Vapnik-Chervonenkis Theory. It is desirable that the detector is robust against the insertion of chaff. We characterize the robustness of $\delta_{d_\mathcal{A}}$ by deriving the minimum chaff required to have nonzero miss probability.

A. Error Probabilities

In this section, we characterize the error probabilities of the detector $\delta_{d_\mathcal{A}}$ as a function of the sample size n, the threshold value ϵ, and the searching class $\mathcal{A}$. It is known that each class of measurable sets is associated with a positive integer called *Vapnik-Chervonenkis dimension (VC-dimension)* which measures the complexity of the class [14]. For a collection $\mathcal{A}$ with finite VC-dimension, we derive the following exponential upper bounds on the error probabilities of $\delta_{d_\mathcal{A}}$.

Theorem 4.1: Let $S = \{(x_i,\, y_i)\}_{i=1}^n$ be drawn i.i.d. from P_{XY}, and $\mathcal{A}$ have finite VC-dimension d. Then for arbitrary distribution P_{XY}, the false alarm probability of $\delta_{d_\mathcal{A}}$ satisfies

$$P_F(\delta_{d_\mathcal{A}}) \leq 8(2n+1)^d e^{-n\epsilon^2/32}.$$

Moreover, if $d_\mathcal{A}(P_X \circ P_Y, P_{XY}) > \epsilon$, then the miss probability satisfies

$$P_M(\delta_{d_\mathcal{A}}) \leq 8(2n+1)^d e^{-n(d_\mathcal{A}(P_X \circ P_Y,\, P_{XY})-\epsilon)^2/32}.$$

Proof: See Appendix. ■

Remark: Theorem 4.1 provides uniform upper bounds on the error probabilities of $\delta_{d_\mathcal{A}}$. It guarantees that under any distribution, $\delta_{d_\mathcal{A}}$ can perform arbitrarily well with sufficiently large samples (note that a condition needs to be satisfied for diminishing miss probability). The error exponent for false alarm probability increases with ϵ, whereas that for miss probability decreases with ϵ. Therefore, the threshold ϵ represents a tradeoff between false alarm and miss detection.

B. Robustness Against Chaff

It is shown in [16] that it is possible for the attacker to evade any detector by inserting sufficient chaff. There is, however, a limit on the minimum amount of chaff needed to do so. Specifically, it is shown in [16] that the minimum asymptotic fraction of chaff needed to mimic independent Poisson processes of rates no more than λ is $1/(1 + \lambda\Delta)$ for attacking traffic with bounded delay Δ, and $1/(1 + M)$ for attacking traffic through a host with bounded memory M. This minimum fraction gives fundamental limit on the amount of chaff that any detector can handle.

In this section, we will show that the $\mathcal{A}$-distance detector can achieve robustness arbitrarily close to the fundamental limit for a class of joint distributions called the bivariate exponential distribution, derived by Marshall and Olkin in [17]. A pair of nonnegative random variables $(X,\, Y)$ satisfies the bivariate exponential distribution BVE$(\lambda_1,\, \lambda_2,\, \lambda_{12})$ if its distribution function is given by

$$\Pr\{X > s, Y > t\} = e^{-\lambda_1 s - \lambda_2 t - \lambda_{12} \max(s,\, t)}, \quad s, t > 0. \quad (3)$$

The importance of this definition of bivariate exponential distribution is that it preserves the memoryless property of the univariate exponential distribution.

For the bivariate exponential distribution defined above, we characterize the amount of chaff required to evade the $\mathcal{A}$-distance detector in the following theorem.

Theorem 4.2: Suppose we use the $\mathcal{A}$-distance detector with threshold $\epsilon \in (0,\, 1)$ and $\mathcal{A}$ being the class of 45° bands. If $(S_1,\, S_2)$ is a stepping-stone pair in which the pairs of interarrival times $(X_i,\, Y_i)$ ($i = 1,\, 2, \ldots$) have *i.i.d.* bivariate exponential distribution, and the rates of S_1 and S_2 are bounded by λ, then the minimum fraction of chaff to have nonzero miss probability is lower bounded by $(1-\epsilon)/(1+ M)$ for stepping-stone pairs with bounded memory M, and $(1 - \epsilon)/(1 + \lambda\Delta)$ for stepping-stone pairs with bounded delay Δ.

Proof: See Appendix. ■

Remark: Theorem 4.2 says that the $\mathcal{A}$-distance detector can detect any correlation in bivariate exponential distribution. By Theorem 4.1, we see that by increasing sample size, ϵ can be made arbitrarily close to 0 while keeping the false alarm probability bounded by certain level. Therefore, for long connections, the robustness of the $\mathcal{A}$-distance detector can be arbitrarily close to the optimal.

For the attacker, the actual value of ϵ may be unknown. Then the attacker is faced with a tradeoff between the amount of chaff and the level of protection; he can save $100\epsilon\%$ of chaff by taking the risk of having ϵ correlation.

V. Simulation

We implement the $\mathcal{A}$-distance detector using STB to verify the performance. We let P_{XY} be the bivariate exponential distribution defined in Section IV-B. It is shown in [17] that the correlation coefficient ρ between bivariate exponential random variables X and Y is

$$\rho = \lambda_{12}/(\lambda_1 + \lambda_2 + \lambda_{12}),$$

where λ_i ($i = 1,\ 2,\ 12$) are parameters in the definition (3). We will test the performance of the $\mathcal{A}$-distance detector on processes with bivariate exponentially distributed interarrival times of various correlation levels. In practice, this corresponds to the case when attacking packets arrive according to a Poisson process of rate λ_{12}, and are relayed immediately without delay, but the attacker inserts chaff packets according to independent Poisson processes of rates λ_1 and λ_2 in the incoming and outgoing streams, respectively.

Before starting the simulation, we have to solve a couple of implementation problems. The first problem is how to decide the detection threshold ϵ. In the Neyman-Pearson framework, we want to set the threshold to the smallest possible value as long as the false alarm probability is bounded by a prescribed value $\alpha \in (0,\ 1)$. A common way of setting threshold in nonparametric detection is to use training. Training is computation intensive. Furthermore, the training data is not guaranteed to represent all the normal traffic in a network with many different traffic types. We propose to set the threshold by making the false alarm upper bound in Theorem 4.1 equal to α. Then we can write the threshold as

$$\epsilon(n) = \sqrt{-\frac{32}{n}\log\frac{\alpha}{8(2n+1)^d}},$$

where d is the VC-dimension of $\mathcal{A}$. Theorem 4.1 guarantees that the false alarm probability will be bounded by α under arbitrary interarrival distributions. For the class of 45° bands, it is easy to show by the method of Wenocur and Dudley [18] that $d = 2$.

Next, we need to choose the sample size. Since the threshold $\epsilon(n)$ is conservative, the detector often needs a large number of samples to have reasonably small miss probability. We need a guideline on approximately how many samples are needed to have reasonable detection performance. We use the results in Theorem 4.1 to estimate the minimum sample size. In Theorem 4.1, it is proved that the miss probability decays exponentially fast if $d_{\mathcal{A}}(P_X \circ P_Y, P_{XY}) > \epsilon$. Thus we estimate the minimum sample size as the smallest integer n that satisfies $\epsilon(n) < d_{\mathcal{A}}(P_X \circ P_Y,\ P_{XY})$.

In our simulation, we let $\alpha = 0.1$, and vary the correlation ρ among 0.85, 0.90, 0.95, and 0.99. The simulated miss detection probabilities of STB are plotted in Fig. 2. We see that there is a critical sample size beyond which the miss probability quickly drops from 1 to close to 0, and this critical sample size decreases as the correlation value increases. For $\rho_1 = 0.85$, $\rho_2 = 0.90$, $\rho_3 = 0.95$, and $\rho_4 = 0.99$, our estimates of the minimum sample sizes are $n_1 = 854$, $n_2 = 752$, $n_3 = 666$, and $n_4 = 607$ respectively (see Fig. 2). We see that our estimates agree with the simulation curves very well.

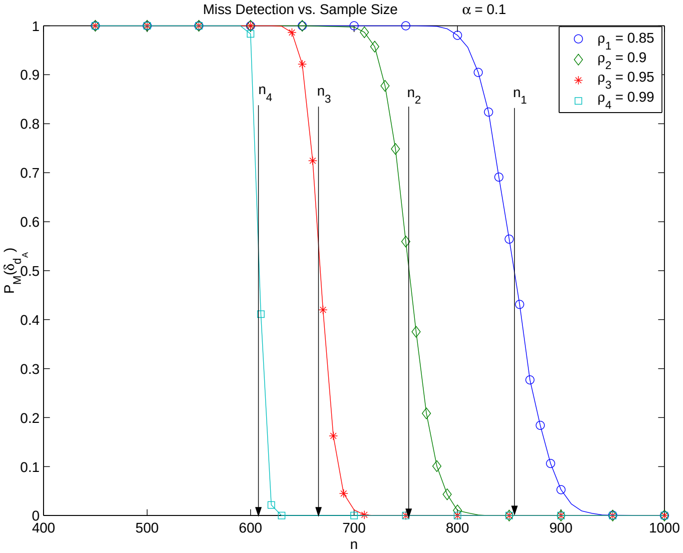

Fig. 2. Simulated miss detection probabilities of STB: $\alpha = 0.1$; 10000 Monte Carlo runs; ρ_i ($i = 1, \dots, 4$): the correlation between X_i and Y_i; n_i: the estimated minimum sample size for ρ_i.

VI. Conclusion

In this paper, we have developed a nonparametric method to detect stepping-stone traffic by correlating the time intervals between packet arrivals. We point out that the *i.i.d.* assumption on pairs of interarrival times is crucial for the proposed detector to work. It means that not only do the processes in consideration need to be renewal marginally, but their pair has to be renewal as well. In practice, this detector should be combined with a preprocessor to filter out the non-renewal processes.

VII. Appendix

A. Proof of Theorem 4.1

The proof uses results derived from the Vapnik-Chervonenkis Theory. In [15], we have proved that for arbitrary distribution P, if S is a collection of n i.i.d. samples drawn from P, and $\mathcal{A}$ is a class of measurable sets with VC-dimension d, then

$$\Pr\{d_{\mathcal{A}}(S,\ P) > \epsilon\} \le 4(2n+1)^d e^{-n\epsilon^2/8}, \tag{4}$$

where $d_{\mathcal{A}}(S,\, P)$ is the $\mathcal{A}$-distance between the empirical distribution according to S and P. Applying (4), we have

$$\begin{aligned}\Pr\{d_{\mathcal{A}}(S_{XY},\, P_{XY}) > \epsilon\} &\leq 4(2n+1)^d e^{-n\epsilon^2/8}, &(5)\\ \Pr\{d_{\mathcal{A}}(S_X \circ S_Y,\, P_X \circ P_Y) > \epsilon\} &\leq 4(2n+1)^d e^{-n\epsilon^2/8}. &(6)\end{aligned}$$

Now we are ready to bound the error probabilities. Since $d_{\mathcal{A}}(\cdot,\cdot)$ satisfies triangle inequality, we have

$$\begin{aligned}d_{\mathcal{A}}(S_X \circ S_Y,\, S_{XY}) &\leq d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY})\\ &\quad +d_{\mathcal{A}}(S_X \circ S_Y,\, P_X \circ P_Y)\\ &\quad +d_{\mathcal{A}}(S_{XY},\, P_{XY}), &(7)\\ d_{\mathcal{A}}(S_X \circ S_Y,\, S_{XY}) &\geq d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY})\\ &\quad -d_{\mathcal{A}}(S_X \circ S_Y,\, P_X \circ P_Y)\\ &\quad -d_{\mathcal{A}}(S_{XY},\, P_{XY}). &(8)\end{aligned}$$

Under $\mathcal{H}_0$, $d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY}) = 0$. Thus, by (7),

$$\begin{aligned}P_F(\delta_{d_{\mathcal{A}}}) &= \Pr\{d_{\mathcal{A}}(S_X \circ S_Y,\, S_{XY}) > \epsilon\}\\ &\leq \Pr\{d_{\mathcal{A}}(S_X \circ S_Y,\, P_X \circ P_Y)\\ &\qquad +d_{\mathcal{A}}(S_{XY},\, P_{XY}) > \epsilon\} &(9)\\ &\leq \Pr\{d_{\mathcal{A}}(S_X \circ S_Y,\, P_X \circ P_Y) > \frac{\epsilon}{2}\}\\ &\quad +\Pr\{d_{\mathcal{A}}(S_{XY},\, P_{XY}) > \frac{\epsilon}{2}\}\\ &\leq 8(2n+1)^d e^{-n\epsilon^2/32}, &(10)\end{aligned}$$

where (10) is obtained by plugging in (5,6).

Under $\mathcal{H}_1$, if P_{XY} satisfies the condition $d_{\mathcal{A}}(P_X \circ\ P_Y,\, P_{XY}) > \epsilon$, then, by (8), we have

$$\begin{aligned}P_M(\delta_{d_{\mathcal{A}}}) &= \Pr\{d_{\mathcal{A}}(S_X \circ S_Y,\, S_{XY}) \leq \epsilon\}\\ &\leq \Pr\{d_{\mathcal{A}}(S_X \circ S_Y,\, P_X \circ P_Y) + d_{\mathcal{A}}(S_{XY},\, P_{XY})\\ &\qquad\qquad \geq d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY}) - \epsilon\}.\end{aligned}$$

Following the same derivation as after (9) yields

$$P_M(\delta_{d_{\mathcal{A}}}) \leq 8(2n+1)^d e^{-n(d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY})-\epsilon)^2/32}.$$

■

B. Proof of Theorem 4.2

By Theorem 4.1, we see that to have non-vanishing miss probability, the attacker has to make $d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY}) \leq \epsilon$. If P_{XY} is the bivariate exponential distribution (BVE) defined in Section IV-B with correlation ρ, then it is shown in [17] that P_{XY} satisfies $P_{XY}(X = Y) = \rho$. For $\mathcal{A}$ being the class of $45°$ bands, we have that $d_{\mathcal{A}}(P_X \circ P_Y,\, P_{XY}) = P_{XY}(X = Y)$. Thus to evade the $\mathcal{A}$-distance detector, the attacker needs to mimic Poisson processes with correlation $\rho \leq \epsilon$.

In [8], Blum *et al.* present an optimal algorithm called "BOUNDED-GREEDY-MATCH" (BGM) to embed traffic with bounded delay into a pair of arbitrary processes; they show that BGM is optimal in that it always inserts the minimum number of chaff packets. In [16], we propose another algorithm called "BOUNDED-MEMORY-RELAY" (BMR), which inserts the minimum number of chaff packets in embedding traffic through a host with bounded memory into arbitrary processes. Therefore, the best way of making attacking traffic with bounded delay or memory mimic given $(S_1,\, S_2)$ is to embed packet transmissions by BGM or BMR, respectively.

The rest of the proof directly follows from the performance of BGM and BMR. It is shown in [16] that the minimum fractions of chaff inserted by BGM and BMR into a pair of independent Poisson processes of rate bounded by λ are $1/(1+\lambda\Delta)$ and $1/(1+M)$, respectively. Furthermore, for BVE distributions, it is shown in [17] that S_i $(i = 1,\, 2)$ can be written as a superposition of Poisson processes P_i and P_3, where P_1, P_2, and P_3 are independent, with rates λ_1, λ_2, and λ_{12}, respectively. If we embed packets into $(P_1,\, P_2)$ by BGM or BMR (assume P_3 does not contain any chaff), we obtain a lower bound on the fraction of chaff as $(1-\rho)/(1+\lambda\Delta)$ and $(1-\rho)/(1+M)$. Combining these results with the constraint $\rho \leq \epsilon$ completes the proof.

■

References

[1] S. Staniford-Chen and L. Heberlein, "Holding intruders accountable on the internet," in *Proc. the 1995 IEEE Symposium on Security and Privacy*, (Oakland, CA), pp. 39–49, May 1995.

[2] X. Wang, D. Reeves, S. Wu, and J. Yuill, "Sleepy watermark tracing: An active network-based intrusion response framework," in *Proc. of the 16th International Information Security Conference*, pp. 369–384, 2001.

[3] Y. Zhang and V. Paxson, "Detecting stepping stones," in *Proc. the 9th USENIX Security Symposium*, pp. 171–184, August 2000.

[4] K. Yoda and H. Etoh, "Finding a connection chain for tracing intruders," in *6th European Symposium on Research in Computer Security, Lecture Notes in Computer Science 1895*, (Toulouse, France), October 2000.

[5] X. Wang, D. Reeves, and S. Wu, "Inter-packet delay-based correlation for tracing encrypted connections through stepping stones," in *7th European Symposium on Research in Computer Security, Lecture Notes in Computer Science 2502*, pp. 244–263, 2002.

[6] D. Donoho, A. Flesia, U. Shankar, V. Paxson, J. Coit, and S. Staniford, "Multiscale stepping-stone detection: Detecting pairs of jittered interactive streams by exploiting maximum tolerable delay," in *5th International Symposium on Recent Advances in Intrusion Detection, Lecture Notes in Computer Science 2516*, 2002.

[7] X. Wang and D. Reeves, "Robust correlation of encrypted attack traffic through stepping stones by manipulation of inter-packet delays," in *Proc. of the 2003 ACM Conference on Computer and Communications Security*, pp. 20–29, 2003.

[8] A. Blum, D. Song, and S. Venkataraman, "Detection of Interactive Stepping Stones: Algorithms and Confidence Bounds," in *Conference of Recent Advance in Intrusion Detection (RAID)*, (Sophia Antipolis, French Riviera, France), September 2004.

[9] P. Peng, P. Ning, D. Reeves, and X. Wang, “Active Timing-Based Correlation of Perturbed Traffic Flows with Chaff Packets,” in *Proc. 25th IEEE International Conference on Distributed Computing Systems Workshops*, (Columbus, OH), pp. 107–113, June 2005.

[10] L. Zhang, A. Persaud, A. Johson, and Y. Guan, “Detection of Stepping Stone Attack under Delay and Chaff Perturbations,” in *Proc. of the 25th IEEE International Performance Computing and Communications Conference (IPCCC 2006)*, (Phoenix, AZ), April 2006.

[11] T. He and L. Tong, “Detecting Encrypted Stepping-stone Connections.” accepted to IEEE Trans. on Signal Processing, 2006.

[12] S. Kullback, *Information Theory and Statistics*. Wiley, 1959.

[13] T. He, S. Ben-David, and L. Tong, “Nonparametric Change Detection and Estimation in Large Scale Sensor Networks,” *IEEE Transactions on Signal Processing*, vol. 54, pp. 1204–1217, April 2006.

[14] V. Vapnik and A. Y. Chervonenkis, “On the uniform convergence of relative frequencie of events to their probabilities,” *Theory of Probability and its Applications*, vol. 16, pp. 264–280, 1971.

[15] T. He and L. Tong, “On $\mathcal{A}$-distance and Relative $\mathcal{A}$-distance,” Tech. Rep. ACSP-TR-08-04-02, Cornell University, August 2004. `http://acsp.ece.cornell.edu/pubR.html`.

[16] T. He and L. Tong, “Detecting Traffic Flows in Chaff: Fundamental Limits and Robust Algorithms.” submitted to IEEE Trans. on Information Theory, 2006.

[17] A. Marshall and I. Olkin, “A Multivariate Exponential Distribution,” *Journal of the American Statistical Association*, vol. 62, pp. 30–44, Mar. 1967.

[18] R. S. Wenocur and R. M. Dudley, “Some Special Vapnik-Chervonenkis Classes,” *Discrete Mathematics*, vol. 33, pp. 313–318, 1981.