Payoff Based IDS Evaluation

Michael Collins
RedJack, LLC

Abstract

IDS are regularly evaluated by comparing their false positive and false negative rates on ROC curves. However, this mechanism generally ignores both the context within which the IDS operates and the attacker's own ability to adapt to IDS behavior. In this paper, we propose an alternative strategy for evaluating IDS based around multiple strategies. Each strategy defines how an attacker profits from attacking a target, and describes victory conditions for the attacker and defender. By mapping the results of ROC analysis to these strategies, we produce results which evaluate defensive mechanisms by their capacity to frustrate an attacker.

1 Introduction

Since the original LARIAT IDS comparisons, the standard practice for evaluating IDS efficiency has been to compare ROC curves [8]. ROC curves describe the efficacy of an IDS via false positive (FPR) and false negative (FNR) rates, which are described as a function of a third *operating characteristic* and then compared using a variety of metrics such as the equal error rate or the area under the curve.

The use of ROC curves for IDS evaluation has been criticized since the original LARIAT work [10, 1]. Of particular note is Axelsson's analysis of the *base-rate fallacy*, which notes that even if a false positive rate is subjectively small, when a test is executed continuously, the number of actual alarms can be intolerably high. This argument was also extended by Gates and Taylor [7] who argue that the false positive rates given in the existing literature are generally far higher than will be accepted by IDS operators.

We argue that, in the IDS domain, the primary weakness of ROC analysis is the lack of real, dimensional, values for comparison. Since the FNR and FPR are dimensionless, they have no relation to operational concerns such as the number of alerts. In particular, researchers have historically been unable to determine whether our defensive mechanisms actually *inhibit* attackers — a scan detector may have a low false positive rate, but if it can't detect the attack until after the entire network is scanned, it serves little purpose.

The technical contribution of this work is an alternative approach to evaluating IDS efficacy that extends ROC analysis into a profit and loss model. In particular, our mechanism treats the IDS as a design specification for the attacker — as a rational entity, the attacker has goals in attacking a network, and we attempt to model those goals in such a fashion that we can determine whether defenses will frustrate an attacker.

To do so, we model different forms of attacks as zero-sum games between attackers and defenders. For each game, the attacker has a specific strategy and payoff. By calculating a fixed probability of detection (a function of the FNR in ROC analysis), we can evaluate, for example, the likelihood that an attacker can successfully DDoS a target with a botnet of particular size. We demonstrate this mechanism using a simple DDoS scenario and a hypothetical detection function. Using our approach, we estimate how many bots an attacker would need to successfully conduct an attack, and given sufficient defender vigilance, how many additional bots would need to be committed, or how long the attack would succeed. As an exploratory work, we emphasize that the numbers expressed in these tests should be considered qualitative, rather than quantitative — the results provide a foundation for comparison between IDS rather than tolerances.

The remainder of this paper is structured as follows:§2 is a survey of previous work in this field. §3 describes our evaluation mechanism. §4 describes the four strategies presented in this paper and their derivation. §5 provides an example scenario using DDoS. §6 concludes the work.

Strategy	Payoff Unit	k-Round Payoff	Limit	Examples
Acquisition	Bots	$k\mathcal{P}(a)(1-\mathcal{D}(a))^k$	Network Size	Scan-to-own
Reconnaissance	Unknown sites	$k\mathcal{P}(a)\sum_{i=0}^{k}(1-\mathcal{D}(a))^i$	Network Size	Scan-to-identify
Saturation	Bandwidth	$\lvert H\rvert(1-\sum_{i=1}^{k}\mathcal{D}(a)^i)$	Overwhelm Defender	DDoS
Backchannel	Duration	$k\mathcal{P}(a)(1-\mathcal{D}(a))^k$	Reach Transmission Limit	Exfiltration

Table 1: Evaluation Games, Payoffs and Example Attacks

2 Previous Work

Lippmann *et al.*'s work on IDS evaluation [8] is the standard for IDS evaluation, and established the practice of using ROC curves to evaluate IDS efficacy. A general history of ROC analysis, explaining the choice and weakness of the methodology was given by McHugh [10]. Of particular relevance is Axelsson's analysis of False Positive Rates and the base-rate fallacy [1].

This paper refines ideas we have proposed in the past about IDS evaluation [4]. This paper extends upon our previous work to reformulate anomaly detection mechanisms in a paradigm of tolerances and payoffs. Cost-based models have also been proposed by Cárdenas *et al.* [3], Stolfo *et al.* [12] and Gaffney and Ulvila [5]. Our approach differs from these in treating the attacker as a rational entity and evaluating his approach in that context.

Other game-theoretic approaches include Cai *et al.* [2] use of game-theoretic model to honeynets, evaluating various strategies for lying to an attacker in order to draw out the maximum intelligence. Our approach differs in intent, in that we are focused on converting ROC results into some form of concrete value, while Cai *et al.*'s work is focused on quantifying an attacker's skepticism about a honeynet. More germane is Marchang and Tripathi's [9] model of IDS activity in MANETs. However, their work is focused on when an IDS should be active (a concern in MANETs), whereas our work focuses on IDS response to different attacker methods.

3 Evaluation Process

In this section, we describe our methodology for evaluating IDS. To do so, we evaluate the *payoff* an attacker receives over an *observable attack space*. The payoff is the utility that the attacker receives for conducting a particular attack, and the observable attack space is the space of behaviors that an IDS can observe which describe the attack. We compare IDS by examining the payoff that an attacker gets over the same OAS in the same time — an IDS which provides a lower payoff to the attacker should be preferable.

Since our approach models attacks as a zero-sum game, it is worth noting that an attacker could theoretically use the same approach to optimize his own attacks. However, in order to do so, the attacker would need to know the FNR and FPR of the defender's IDS.

The remainder of this section describes the construction of OAS and the problem of timing. §3.1 describes the observable attack space and how we model reaction time. §3.2 addresses the problem of timing and the false positive rate. §3.3 describes payoff and probability of detection.

3.1 Observable Attack Space

An IDS analyzes some form of log data, such as service logs (as is the case with host-based IDS) or captured packets (as is the case with Bro [11], Snort[1] or NetFlow analysis systems [6]). The choice of log data limits the inferences that an IDS can make. For example, since flow analysis systems have no access to payload, they cannot determine with certainty whether a packet is crafted to use a particular buffer overflow or a password attack.

To model the limitations of log analysis systems, we refer to the space of attacker activity *as observed by the IDS* as the *Observable Attack Space* (OAS). The OAS is a multidimensional space of attributes derived from traffic log data. The OAS relates the attacker's payoff to the defensive mechanism being used. As a result, OAS will vary as a function of the log data collected, the attack strategy (see §4) and the IDS.

The dimensions of the OAS are a subset of the attributes found within the log data, and may be aggregates or permutations of those attributes. At the minimum, the OAS' dimensions should consist of whatever attributes the attacker uses to determine payoff and whatever attributes the IDS uses to infer behavior. For example, NetFlow data contains among its fields the source and destination IP addresses of a flow, the protocol used and the time that the flow occurred. An OAS generated from flow data may use the number of unique flows observed per IP address, or the number of bytes and packets observed. An IDS which is monitoring scanning may be evaluated using an OAS whose dimensions consist of the number of IP addresses contacted by a single IP, and the number of addresses where a legitimate session was recorded.

3.2 Reaction Time And False Positive Rate

We now address the problem of timing. An IDS requires some time to respond to an attack. To simplify evaluation, we model response time using discrete periods. In this model, attacks take place in *rounds*: discrete, fixed-length periods after each of which the defensive system can judge traffic and respond to attacker behavior.

The defender chooses the length of the round; however, round length bounds the FPR. We assume that any IDS is being manned by some operator who judges or validates the results. Given this constraint, we set an acceptable limit of one false positive per 8 hour operator shift, and then calculate the FPR correspondingly. For example, if the round length is one minute, then the maximum allowable FPR is 0.2%.

3.3 Detection Probability and Payoff

Once an OAS and round length have been determined, we map attacks to the OAS as discrete *attack points*. For any attack point, a, we define the *payoff* $\mathcal{P}(a)$ as the utility to the attacker of conducting that particular attack. The dimensions of the payoff vary based on the strategy the attacker executes and are discussed in §3. Each strategy defines attacker utility differently, and can include acquiring bots, gaining intelligence about an unknown network, or simply keeping a connection open as long as possible.

The complement to payoff is *probability of detection*. $\mathcal{D}(a)$ is the probability that a defender detects the attacker, and responds at the end of a round. The defender estimates the FNR rate using a ROC curve, and then calculates the probability of detection $1 - \mathsf{FNR}$. Recall that the round length fixes the FPR to a small value; we treat the FNR as fixed at its highest value within that range (*i.e.*, if the FPR is 0.001 and the FNR ranges between 0.1 and 0.3, we treat it as 0.3 and use a corresponding $\mathcal{D}$ of 0.7).

4 Strategies

A *strategy* is a model of an attack consisting of a *payoff function* and a *limit*. The payoff function is the attacker's gain over a sequence of discrete rounds, and the limit is the aggregate payoff defining attacker victory. For each strategy x, we define an aggregate payoff function $\mathcal{A}_{\mathrm{x}}(\mathcal{D}(a), k)$, which is the total payoff after k rounds. The attacker wins in any round where $\mathcal{A}_{\mathrm{x}}$ exceeds the limit, and may depending on the strategy in question, opt at that point to quit. For example, when scanning a network, the attacker's payoff is the number of hosts contacted and the limit is the number of hosts on the network. After scanning the entire network, the attacker quits.

The aggregate payoff function is defined by the attacker and the defender's interaction. Defenders react to each attack slightly differently, but we assume that defenders have a limited set of options: they can only block attacking hosts or restore a compromised host on their network. Defenders have no impact on any network except their own.

In this paper, we define four strategies, which are summarized in Table 1. The four strategies are *acquisition* (the takeover of a host via scanning), *reconnaissance* (simple scanning), *saturation* (DDoS), and *backchannel communications* (covert channel communication or DDoS from the attacking side). Table 1 summarizes these strategies, an example payoff unit, their aggregate function and limits. The strategies described in this document are simple, and are not intended as a comprehensive set, but as an initial collection of descriptors.

4.1 Acquisition

The *acquisition* strategy models the takeover of multiple remote hosts by an attacker using some form of exploit, such as a buffer overflow or a password list. In this strategy, the attacker controls a single host which communicates with the defender's network. In each round, he communicates with some number of hosts in the targeted network, and receives a payoff for each host he successfully contacts.

In the acquisition strategy, the defender will block the attacker if detected, and restore any hosts the attacker compromised. As a result, the aggregate payoff for acquisition is a function of whether, in the k rounds the attacker operates, the defender *ever* detects him. If detected, the attacker's payoff is zero, if undetected the payoff is $k\mathcal{P}(a)$. Equation 1 formalizes this value.

$$\mathcal{A}_{\mathrm{acq}} = k\mathcal{P}(a)(1 - \mathcal{D}(a))^k \tag{1}$$

The limit for acquisition is the size of the network the defender is protecting.

4.2 Reconnaissance

The *reconnaissance* strategy models scanning and other attempts to gather intelligence on a network's structure. Reconnaissance differs from acquisition in that the attacker is interested only in communicating with hosts, not compromising them. Because of this, the attacker does not modify the network and all the defender can do to impact payoff is block further communications. If detected, the attacker keeps the payoff from previous rounds. As such, for k rounds, the aggregate payoff for

reconnaissance is simply a function of whether the defender was detected:

$$\mathcal{A}_{\text{rec}} = k\mathcal{P}(a)\sum_{i=0}^{k}(1-\mathcal{D}(a))^i \quad (2)$$

Note that since the attacker can only be blocked, it gets the first round (before the defender has a chance to react). The reconnaissance strategy succeeds if the attacker is able to communicate with the entire observed network without being detected.

4.3 Saturation

The *saturation* strategy models DDoS and other mechanisms where the attacker's goal is to swamp a connection with traffic. In saturation, the attacker begins with a set of hosts, H. In each round, each host attacks at attack point a, and the attacker payoff is the product of his payoff for an individual host across the entire set.

In each round of the saturation strategy, the defender identifies $|H| \cdot \mathcal{D}(a)$ hosts and progressively blocks them at the end of the round. As a result, the k-th round aggregate payoff for the attacker is a function of the number of hosts the defender has been able to block, and is expressed as:

$$\mathcal{A}_{\text{sat}} = |H|(1-\sum_{i=1}^{k}\mathcal{D}(a)^i) \quad (3)$$

The limit for saturation is a function of the target and its tolerance. For example, if the target is a webserver, the limit may be a function of the number of requests that are processed in a round. If the target is a router, the limit may be the bandwidth of an interface on the router. In addition, unlike the acquisition and reconnaissance strategies, saturation can continue indefinitely — reaching the limit in saturation is an indicator that the attacker is able to block communications for the next round.

4.4 Backchannel

In the *backchannel* strategy, the attacker owns a compromised host within the defender's network and is using it for its own purposes. Backchannel communications may include dialing home to a botnet command and control server, or downloading and uploading files.

In the backchannel scenario, the attacker's goal is to maintain sufficient control of the host to transfer a file of specific size (the limit in the scenario) without being detected. If the communication is detected, then the defender identifies the host and restores it, yielding an effective payoff of zero. The aggregate payoff is therefore identical to the acquisition scenario in that if the attacker is *ever* detected, the payoff is zero:

$$\mathcal{A}_{\text{bkc}} = k\mathcal{P}(a)(1-\mathcal{D}(a))^k \quad (4)$$

5 Evaluating an IDS

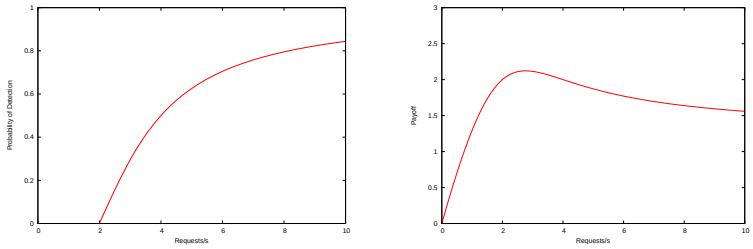
In this section, we apply our methodology to explore attack and defense options during a DDoS. We stress that the models of traffic used in this section are synthetic — the focus of the work in this paper is not on developing a defensive mechanism, but evaluating how those mechanisms can be applied. Our approach works evaluating the impact that defensive mechanisms have on attacker goals. When evaluating a single system, we evaluate how different attacker choices (represented as different attack points) impact his goal of reaching the limit for whatever strategy he is applying.

For brevity, we demonstrate how a single defense can be used to evaluate its impact on attacks; our earlier work [4] gives an example of comparison. We consider a simple DDoS scenario where an attacker controls a small botnet consisting of 10,000 bots. We consider DDoS because it can be modeled as an endurance contest — the defender wins if he can continue to operate in the presence of the attacker's attacks. For our model, the defender is defending a single HTTP server with a capacity of 500 requests/s[2]. To evaluate the impact of an IDS on an attack, we must define an OAS, generate $\mathcal{D}$ and $\mathcal{P}$, and then evaluate their impact using Equation 3.

For the purposes of this example, we choose a simple 1-dimensional OAS — a is the number of requests originating from the host in a 1 second round. We model attacker payoff as $\mathcal{P}(a) = a$; that is, the attacker's payoff is the number of requests fired off by a host in a second. We model the probability of detection, $\mathcal{D}(a)$ as: $\mathcal{D}(a) = \frac{\tan^{-1}((a-2)/2)}{\pi/2}$. This function is chosen to yield the curve in Figure 1(a), which tends asymptotically towards 1 as a (measured in requests/s) tends towards ∞. Figure 1(b) shows the expected 1-round, 1-host payoff ($\mathcal{D}(a) \cdot \mathcal{P}(a)$) for an attack as a function of the number of requests/s generated by the bots. We get a maximum integral payoff if an attacking host sends 3 requests/s.

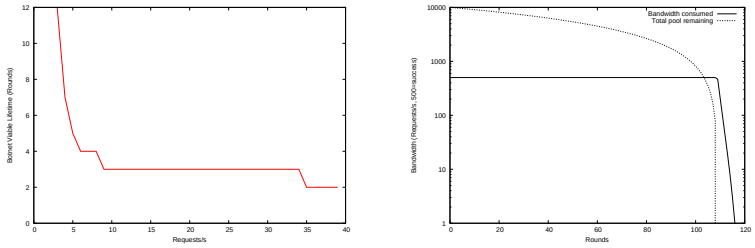
Using equation 3, we can now evaluate various approaches to DDoS by the attacker. Figure 2(a) calculates the effective lifetime of the 10,000 host botnet by assuming that the attacker commits all 10,000 hosts initially and then applying Equation 3 directly. Figure 2(a) calculates the largest value of k for which, given $|H| = 10000$, the aggregate payoff is greater than 500 Requests/s. The longest lifetimes are achieved by subtlety, and the attacker gains more by conducting short requests.

An alternate scenario is one where the attacker replaces hosts as they are blocked. The efficiency of this approach is shown in Figure 2(b), which plots the ex-



(a) Probability of Detection (b) Weighted Payoff

Figure 1: Metrics describing the performance of the attacker botnet in the DDoS scenario.



(a) Lifetime with total commitment (b) Lifetime with ideal replacement

Figure 2: Expected botnet lifetimes using different replacement strategies.

haustion of his pool of bots as he conducts this attack. In this case, we use the optimal individual payoff from Figure 1(b), 3 Requests/s/host, and a steady stream of 167 hosts. At the end of each round, the attacker adds as many hosts are required by calculating the loss using Equation 3 with $k = 1$.

Finally, we consider the situation where the attacker is unaware of the defensive capacities of the network and will therefore vary his offense. To do so, we plot the attacker's replacement rate as function of the bots committed and their aggressiveness over a space of values. Figure 3 shows an example of this approach. In this case, only successful attacks (ones where the total payoff exceeds the limit) are plotted.

The results of Figures 2 and 3 provide an alternative means of evaluating the efficiency of detection mechanisms. The payoff and limit values can be applied in multiple ways to evaluate defensive efficiency. If a defense prevents the attacker from reaching his limit, then it could be considered objectively successful in that the defender will survive attacks of that form; however, we expect such a situation to be rare. Alternatively, when comparing multiple systems, system A is more successful than system B when it produces a lower payoff than B at the same attack point. Another approach is to evaluate how well an attacker can achieve a particular payoff — that is, how long it takes, or whether a particular payoff can be maintained indefinitely.

6 Conclusions

In this paper, we have introduced an alternative method of evaluating the efficacy of detection and defensive mechanisms based around payoff functions. In comparison to ROC-based evaluation, this approach expands the false positive and false negative rates to dimensional values to measure the impact that a defensive mechanism has on attacker goals.

As exploratory work, the results of this evaluation strategy are rough, and best considered as qualitative, rather than quantitative — they are viable for describing whether one defensive approach is notably better than another, but the resulting numbers are not yet realistically connected with real-world phenomena. Future work will focus on refining these scenarios to more accurately measure real-world behavior such as server exhaustion.

References

[1] S. Axelsson. The base rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Secu-*

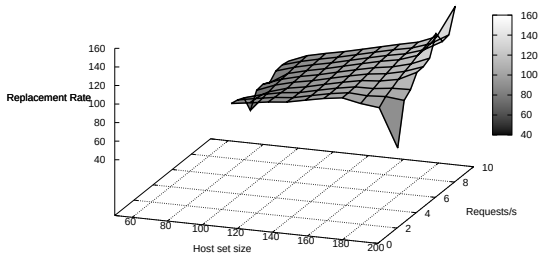


Figure 3: Replacement rate as a function of $|H|$ and a. Only populations exceeding the limit of 500 Requests/s are plotted.

rity, 3(3):186–205, 2000.

[2] J. Cai, V. Yegneswaran, C. Alfeld, and P. Barford. An attacker defender game for honeynets. In *Proceedings of the 2009 COCOON Conference*, 2009.

[3] A. Cárdenas, J. Baras, and K. Seamon. A framework for evaluation of intrusion detection systems. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, 2006.

[4] M. Collins and M. Reiter. On the limits of payload-oblivious network attack detection. In *Proceedings of the 2008 RAID Symposium*, 2008.

[5] J. Gaffney and J. Ulvila. Evaluation of intrusion detectors: A decision theory approach. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, 2001.

[6] C. Gates, M. Collins, M. Duggan, A.Kompanek, and M. Thomas. More netflow tools: For performance and security. In *Proceedings of the 18th Large Installation Systems Administration Conference (LISA 2004)*, 2004.

[7] C. Gates and C. Taylor. Challenging the anomaly detection paradigm: a provocative discussion. In *NSPW '06: Proceedings of the 2006 workshop on New security paradigms*, 2007.

[8] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, and M. Zissman. Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation. In *Proceedings of the DARPA Information Survivability Conference and Exposition*, 2000.

[9] N. Marchang and R. Tripathi. A game theoretical approach for efficient deployment of intrusion detection systems in mobile ad hoc networks. In *Proceedings of the 2007 ADCOM Conference*, 2007.

[10] J. McHugh. Testing intrusion detection systems: a critique of the 1998 and 1998 DARPA intrusion detection system evaluations as performed by lincoln laboratory. *IEEE Transactions on Information and Systems Security*, 3(4):262–294, 2000.

[11] V. Paxson. Bro: A system for detection network intruders in real time. In *Proceedings of the 2008 Usenix Security Symposium*, 1998.

[12] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan. Cost-based modeling for fraud and intrusion detection: Results from the JAM project. In *Proceedings of the 2000 DARPA Information Survivability Conference and Exposition*, 2000.

Notes

[1] http://www.snort.org

[2] Capacity estimate from http://httpd.apache.org/docs/2.2/mod/mod_unique_id.html, fetched May 29,2009