Learning Entropy

Lele Zhang and Darryl Veitch

Department of Electrical and Electronic Engineering
The University of Melbourne, Parkville Victoria 3010, Australia
{lz,dveitch}@unimelb.edu.au

Abstract. Entropy has been widely used for anomaly detection in various disciplines. One such is in network attack detection, where its role is to detect significant changes in underlying distribution shape due to anomalous behaviour such as attacks. In this paper, we point out that entropy has significant blind spots, which can be made use by adversaries to evade detection. To illustrate the potential pitfalls, we give an in-principle analysis of network attack detection, in which we design a camouflage technique and show analytically that it can perfectly mask attacks from entropy based detector with low costs in terms of the volume of traffic brought in for camouflage. Finally, we illustrate and apply our technique to both synthetic distributions and ones taken from real traffic traces, and show how attacks undermine the detector.

Keywords: Entropy, Anomaly Detection, Camouflage

1 Introduction

Entropy is widely used as a summary statistic in diverse application areas, including *anomaly detection*. In recent years anomaly detection has received considerable attention in computer networking both within industry and academia, in particular in relation to security issues such as network based attacks, and entropy based detection is a popular approach.

Generally speaking, in entropy based anomaly detection, the (empirical) entropy, summarising a histogram of some quantity of interest from the underlying network traffic within a time bin, is used as a detector of anomalous events seen across bins. For example, the quantity could be the counts of packets (becoming probabilities after normalisation) with different source port numbers passing a measurement point. Implicitly, it is assumed/believed that entropy will change noticeably when 'significant' changes in the traffic pattern occur due to anomalous behaviours, but change little or not at all when small fluctuations about typical behaviour are encountered. In this vein, [3, 5] used entropy of source IP address distributions to capture DDoS attacks, and [10] focused on worm detection using distributions from packet headers. The paper [7] considered entropy of distributions based on the number IP addresses that each host communicates with in addition to those from packet headers. Other work exploiting entropy for anomaly detection can be found in [1, 4].

These prior works have demonstrated a measure of success of entropy-based anomaly detection, especially in network attack detection. However, this success is founded on two basic assumptions that we challenge here: *i*) Entropy detects 'significant' changes in distribution well, *ii*) Attackers are unable and/or unwilling to adopt strategies to evade detection. In fact, since entropy, like all summary statistics, is a compact summary of a complex reality, it must necessarily be blind to many changes in the underlying distribution. For anomaly detection purposes, it is clearly essential to understand the nature of this blindness, however this task has until now not attracted much attention.

In this paper, we shall see that the relationship between changes in distribution details, or 'shape', and entropy can be complex and counter-intuitive. An important consequence for entropy based detectors is that they can be surprisingly easily defeated when the attacker starts to learn and manipulate data. Sophisticated adversaries have been discussed before in the context of detectors (or classifiers) in other domains [2, 6, 8]. Since entropy based detection is becoming popular, we believe that exploring its limitations and potential countermeasures is both important and timely.

Our main contributions: *i*) The first quantitative understanding of the behaviour of entropy as a function of the underlying distribution shape. *ii*) To demonstrate the potential fruits of (*i*), we apply it to the attack detection problem. We provide the definition and first results on optimal camouflage from entropy based detectors.

The above results bring insights and capabilities at a number of (closely related) levels, including the nature of entropy blindness, how attackers can evade detection at minimal cost, and a meaningful calibration capability for detectors and understanding of their limitations. Although it has been noted before that entropy is not all things to all applications, for example see [3, 9, 11], this is the first study we are aware of which provides a rigorous quantitative analysis and a systematic investigation. We believe that the insights are very valuable for more general settings and also that the techniques can be extended to analyse more realistic attack scenarios.

In Section 2 examples are used to illustrate some of the key traps one can fall into from assuming that entropy captures changes in distribution shape. Section 3 develops the technical results that enable 'optimal camouflage', which are then used in Section 4 to explore attack detection using an entropy based detector. We conclude and discuss future work in Section 5.

2 Entropy

After introducing the definition of entropy, we explore, using a number of concrete examples, the dangers of simplistic impressions as to the relationship between distribution shape and entropy. In particular, these examples give insight into what detectors might hope to detect and what they might miss, and hence help us better understand why attackers can evade the entropy detector with surprising ease.

2.1 Preliminaries

We view Internet traffic as an infinite stream of packets passing by a passive monitoring point, processed in batches according to contiguous constant width measurement intervals. Within a single such interval consisting of V packets, and given a metric of interest (a function of packet header information), each packet is mapped to one of N classes, resulting in a sequence $\{a_1, a_2, \ldots, a_V\}$ over the bin where $a_j \in \{1, 2, \ldots, N\}$. The associated empirical distribution $\mathcal{D}$ is given by $\mathcal{D} = \{p_1, p_2, \ldots, p_N\}$, where $p_i = f_i/V$, $f_i = |\{j : a_j = i\}|$ and $\sum_{i=1}^{N} p_i = 1$.

The Shannon (empirical) entropy for $\mathcal{D}$ is defined as

$$\mathcal{H}(\mathcal{D}) = -\sum_{i=1}^{N} p_i \log p_i,$$

where $0 \log 0 = 0$, and logarithms are base 2. Provided the alphabet size N is finite, entropy is maximal when the distribution is uniform $\{1/N, 1/N, \ldots, 1/N\}$. In contrast, the minimum possible entropy of zero occurs when probability is maximally concentrated: $p_j = 1$ for some j and $p_i = 0$ otherwise.

2.2 Connecting Entropy and Distribution Shape

As a statistic summarising a distribution, entropy might be expected to be alike for distributions of similar shape, that is, those with only small differences in their probabilities. On the other hand it would be substantially different for ones with radically different shapes. In fact, implicitly or explicitly, this is one of the principles that entropy based detection relies on. Unfortunately, neither of these hold true as we now show.

Order: Although there is no commonly agreed non-parametric definition of distribution 'shape', few would claim that the distributions, $\mathcal{D}_{G1}$ and $\mathcal{D}_{G2}$, shown in Fig 1(a),(b) have similar shapes. However, they share exactly the same entropy due to the fact that $\mathcal{D}_{G1}$ (a truncated, renormalised, geometric distribution with $p = 0.607$), is an ordered version of $\mathcal{D}_{G2}$ and entropy is *blind to order*. Although this is well understood, nonetheless order invariance gives a wealth of examples where very large probability differences are not reflected at all in entropy.

Shape: A natural question is whether all 'dramatic' examples involve ordering. The answer is no. Unless $N = 2$, the set of distributions with a given $\mathcal{H} > 0$ is uncountably infinite, even if they are first ordered. More concretely, consider the examples in Fig 1(c),(d). One can ask "Which distribution is closer to $\mathcal{D}_{G1}$, $\mathcal{D}_{G3}$ or $\mathcal{D}_v$?", to which "$\mathcal{D}_v$" is probably the answer of 99.9% of readers, including statisticians. Not only do the shapes of $\mathcal{D}_{G1}$ and $\mathcal{D}_v$ seem qualitatively very similar, but the probability ranges $[0.0044, 0.3961]$ and $[0.0092, 0.3358]$ are also quite close, compared to $[0.0493, 0.5560]$ from $\mathcal{D}_{G3}$. However, entropy concludes differently. In fact $\mathcal{D}_{G3}$ has the same entropy as $\mathcal{D}_{G1}$, whereas $\mathcal{D}_v$ has an entropy over 10% greater.

Heavy hitters: We would reasonably expect that entropy would be relatively sensitive to large changes in the most significant probabilities, or 'heavy hitters',

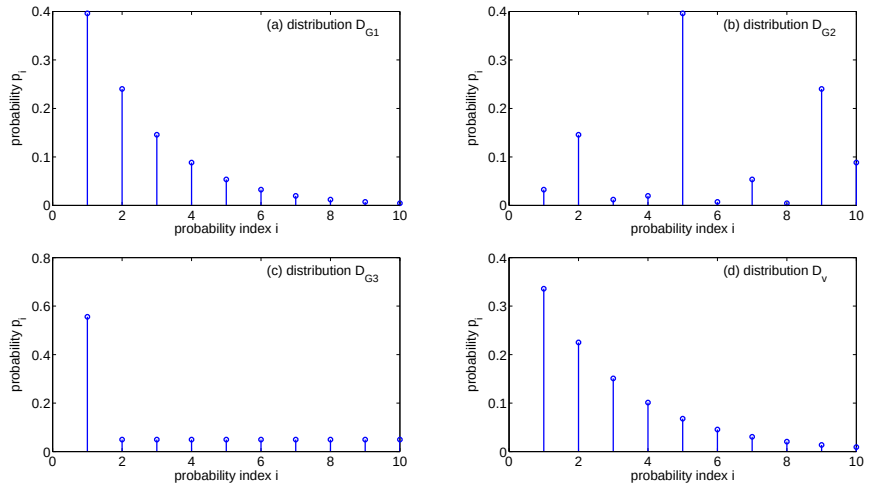

Fig. 1. Four distributions $\mathcal{D}_{G1}, \mathcal{D}_{G2}, \mathcal{D}_{G3}$ and $\mathcal{D}_v$ with $N = 10$ where: $\mathcal{H}(\mathcal{D}_{G1}) = \mathcal{H}(\mathcal{D}_{G2}) = \mathcal{H}(\mathcal{D}_{G3}) = 2.40 < \mathcal{H}(\mathcal{D}_v) = 2.64$.

as low probabilities contribute little individually to $\mathcal{H}$ since $p_i \log p_i \to 0$ when $p_i \to 0$. In fact this can be far from the case.

Suppose that outcome $i = 1$ of $\mathcal{D}_{G1}$ disappears for some reason (such as all traffic from a very popular server being rerouted around the monitoring point). Thus its probability p_1, the largest carrying almost 40% of the total, falls to zero, yet despite this the new renormalised distribution has an entropy of 2.37 compared to 2.40 originally, a mere 1.3% decline. The reason lies in the shape of the remainder of the (reordered) distribution combined with the compensating effect of renormalisation.

More generally, the following expression describes the impact of the removal of a single outcome:

$$\mathcal{H}_0 = -p_l \log p_l - (1 - p_l) \log(1 - p_l) + (1 - p_l)\mathcal{H}(\mathcal{D}_R), \tag{1}$$

where p_l is the missing probability, and $\mathcal{H}(\mathcal{D}_R)$ the entropy of the renormalised distribution $\mathcal{D}_R$ consisting of all the other probabilities except p_l. Clearly there are cases when $\mathcal{H}_0 = \mathcal{H}(\mathcal{D}_R)$, that is when the missing probability has no impact on the entropy at all! In fact, simple algebra shows that the removal of the r largest probabilities of a truncated geometric distribution is the same as that resulting from the removal of the r smallest ones. Thus, failure to track the largest probabilities does not affect entropy much since their absence has the same effect as omitting the same number of small probabilities, which is small (provided r is not too large).

Although the details of the above analysis were based on the special properties of the geometric distribution, it nonetheless highlights a far more general point, that assumptions on how entropy is influenced by distribution features cannot be taken for granted.

It is of course well known that, like other statistics, entropy provides only a coarse summary of data, and so must necessarily be blind to many features. The key message of this section is that the details of the connection between distribution 'shape' and entropy may be counterintuitive, which in the context of anomaly detection strongly motivates a precise quantitative understanding of the connection between the two.

3 Entropy in Attack Detection

In this section we examine the use of entropy in the context of Internet attacks and their measurement based detection. Here an *attacker* sends attack traffic into the network, which he may attempt to disguise, and the *detector* seeks to detect his activities using an entropy based anomaly detector. In this paper our explicit focus is on optimal camouflage strategies for the attacker, but these two sides of the same battle are very closely related as we make clearer below.

Here we flesh out entropy-specific issues using an idealised proof of concept model, rather than attempting to provide a realistic description of network attacks or detection. Our aim is to provide rigorous results and insights, and an approach, which can be used as tools in the study of entropy-based anomaly detection. In Section 4 we give an example showing how more complex scenarios can be built up using the building blocks we provide.

3.1 Model

Detector We assume that the detector knows a benchmark distribution $\mathcal{D}_0$ with an entropy of $\mathcal{H}_0$ corresponding to normal (and attack free) traffic, and also knows the histogram $\mathcal{D}$ from the current time interval, summarised by its empirical entropy $\mathcal{H}$. The detection mechanism is simple: an attack is declared if $\mathcal{H}$ differs from $\mathcal{H}_0$ by more than a threshold $\theta_{\mathcal{H}} \geq 0$, or the detector is silent if $\mathcal{H}$ falls in the interval $[(1-\theta_{\mathcal{H}})\mathcal{H}_0, (1+\theta_{\mathcal{H}})\mathcal{H}_0]$.

In practice there are many issues making the determination of $\mathcal{D}_0$ and $\mathcal{H}_0$ difficult, and furthermore, $\mathcal{H}$ cannot in general be precisely measured. We subsume each of these effects into the need to describe the *sensitivity* of the detector, which we do through $\theta_{\mathcal{H}}$.

Attacker For our purpose an *attack* is the presence of packets sent by the attacker with malicious intent which pass by the monitoring point during the measurement interval. We model it by the number $V_{\mathcal{A}}$ of attack packets, the set $\mathcal{T}$ of class indices in which they appear for the chosen metric, and their distribution across $\mathcal{T}$. During an attack the measurement interval contains $V_0 + V_{\mathcal{A}}$ packets with an *attack intensity* of $V_{\mathcal{A}}$, resulting in the modified distribution $\mathcal{D}$.

Attacks require resources to mount. We measure the cost of an attack (per measurement interval) to the attacker by the number of packets he sends. For a given $V_{\mathcal{A}}$, depending on the nature of the attack may have some flexibility through the choice of $\mathcal{T}$ to reduce his impact on $\mathcal{H}$ and hence the chance of detection. We call this *passive camouflage*. Here we assume that the attacker

knows $\mathcal{D}_0$. Although this is a quite conservative assumption in practice, it is plausible that the attacker could learn it over time. When passive camouflage is impossible or insufficient, then the attacker may opt to augment it by using *active camouflage* through sending a number $V_\mathcal{C}$ of additional *camouflage packets* in a tailored spread over indices design to further reduce the impact on $\mathcal{H}$. So the traffic volume becomes $V_0 + V_\mathcal{A} + V_\mathcal{C}$ with a camouflage cost of $V_\mathcal{C}$. The resulting camouflaged distribution is given by $\mathcal{D}_C$. How to camouflage actively and efficiently is one of the main points we focus on.

3.2 Optimal Camouflage

The problem of designing camouflage strategies is formulated as follows. The attacked traffic histogram $\mathcal{D} = \{p_1, \ldots, p_N\}$ (without loss of generality, indexed in non-increasing order) has entropy $\mathcal{H}$ which by definition is outside of the perceived *normal range* $[(1-\theta_\mathcal{H})\mathcal{H}_0, (1+\theta_\mathcal{H})\mathcal{H}_0]$, and so will trigger an alarm. To hide from the detector, the attacker must disguise itself by 'dragging' the entropy back to some *target entropy* $\mathcal{H}_T$ lying within the normal range. According to the difference between $\mathcal{H}$ and $\mathcal{H}_T$ the strategies are different. In general, if $\mathcal{H}_T > \mathcal{H}$ then the attacker needs to equalise the probabilities; otherwise, he should concentrate them.

Changes in the probabilities can be achieved primarily in two ways: through sending extra packets (targeted increments), or by somehow removing normal traffic (decrements). We consider the 'increment only' scenario in this paper, as this is clearly directly feasible for the attacker.

We consider *optimal camouflage*, that is how to achieve a given $\mathcal{H}_T$ at minimal cost, that is, with the smallest possible number $V_\mathcal{C}$ of camouflage packets sent. This also reduces the chances that the attack would be captured by other techniques, such as by volume detection.

Formulating the original problem: Let δ_i denote the increment of probability p_i due to camouflage for constant V_0 and $V_\mathcal{A}$. Then the camouflage cost is calculated as $V_\mathcal{C} = (V_0 + V_\mathcal{A})\Delta$ with $\Delta = \sum_{i=1}^{N} \delta_i$. Hence the optimisation problem can be formulated as:

$$\mathbf{Min}\Delta: \quad \min_{\{\delta_i\}} \Delta, \quad \text{s.t. } \delta_i \geq 0 \,\forall i \text{ and } \mathcal{H}(\mathcal{D}_C) = \mathcal{H}_T,$$

where $\mathcal{D}_C = \{(p_i + \delta_i)/(1 + \Delta),\, i = 1, 2, \ldots, N\}$ and $1 + \Delta$ renormalises the distribution since all increments are positive.

In other words, we find the smallest increment 'budget' which can achieve the target entropy. Since the objective function Δ is a component of the renomalisation factor, it turns out that this problem is best solved through first solving the inverse problem, where we find the extremal $\mathcal{H}$ using (all of) a fixed budget.

Formulating the inverse problem: If $\mathcal{H}_T > \mathcal{H}$ then the inverse problem is

$$\mathbf{MaxH}: \quad \max_{\{\delta_i\}} \mathcal{H}(\mathcal{D}_C), \quad \text{s.t. } \delta_i \geq 0 \text{ and } \sum_{i=1}^{N} \delta_i = c,$$

for a constant c. Otherwise, if $\mathcal{H}_T < \mathcal{H}$ it becomes

$$\textbf{MinH}: \quad \min_{\{\delta_i\}} \mathcal{H}(\mathcal{D}_C), \quad \text{s.t. } \delta_i \geq 0 \text{ and } \sum_{i=1}^{N} \delta_i = c.$$

As the renormalisation factor, $1+c$, is determined, the above problems are equivalent to the ones below under the same constraints (resp. **MaxH** and **MinH**):

$$\textbf{Min}\mathrm{H}^{-}: \min_{\{\delta_i\}} \sum_{i=1}^{N} (p_i+\delta_i)\log(p_i+\delta_i) \text{ and } \textbf{Max}\mathrm{H}^{-}: \max_{\{\delta_i\}} \sum_{i=1}^{N} (p_i+\delta_i)\log(p_i+\delta_i).$$

Solving the inverse problems: Consider **MaxH**, that is to solve **Min**H^{-}, whose objective function is convex and constraints belong to a convex set. The global minimum can be solved using Lagrange multipliers and the Karush-Kuhn-Tucker (KKT) conditions. We define the Lagrange function as:

$$\Lambda = \sum_{i=1}^{N} (p_i + \delta_i)\log(p_i + \delta_i) + \mu(\sum_{i=1}^{N} \delta_i - c) - \sum_{i=1}^{N} \lambda_i \delta_i.$$

For simplicity, we consider the natural logarithm here. The KKT conditions are given by $\sum_{i=1}^{N} \delta_i - c = 0$, $\log(p_i + \delta_i) + 1 + \mu - \lambda_i = 0$, $\lambda_i \delta_i = 0$, $\delta_i \geq 0$ and $\lambda_i \geq 0$, for all i.

Consider the following two cases: *i*) If $\lambda_i > 0$, then $\delta_i = 0$. Because $\delta_i = e^{\lambda_i - \mu - 1} - p_i > e^{-\mu-1} - p_i$, we can write $\delta_i = (e^{-\mu-1} - p_i)^{+}$. *ii*) If $\lambda_i = 0$, then $\delta_i = e^{-\mu-1} - p_i \geq 0$. Overall, δ_i can be written as $\delta_i = (e^{-\mu-1} - p_i)^{+}$. Then μ is determined by $\sum_{i=1}^{N} (e^{-\mu-1} - p_i)^{+} - c = 0$, followed by the solution for δ_i.

We see that in the optimal solution $\{\delta_i\}$ decomposes naturally into two subsets. One contains zero δ_i's, which are applied to large probabilities that stay invariant. The other consists of positive δ_i's, which are given to small probabilities in order to raise them to a common level, namely $p_i + \delta_i = e^{-\mu-1}$.

To solve **MinH** (i.e. **Max**H^{-}), we make use of the following inequality.

$$\begin{aligned}(p_x + \delta_x)\log(p_x + \delta_x) + (p_y + \delta_y)\log(p_y + \delta_y) \geq \\ (p_x + \delta_x + \delta_y)\log(p_x + \delta_x + \delta_y) + p_y \log p_y\end{aligned} \tag{2}$$

if $p_x \geq p_y \geq 0$ and $\delta_x, \delta_y \geq 0$, which follows the fact that the function $F_\epsilon(z) = (z+\epsilon)\log(z+\epsilon) - z\log z$ is strictly monotonically increasing for any $\epsilon > 0$ and $z > 0$. In fact, Equation (2) states that the difference between two probabilities grows (with all others constant), then the entropy drops, whereas the entropy rises when the contrast between them reduces. Following (2), clearly, the optimal solution for **MinH** is $\delta_1 = c$ and $\delta_i = 0\ \forall i \neq 1$ because moving all increments to the largest probability p_1 reduces the overall entropy.

Solving the original problem: From the solutions to the inverse problems we observe that the maximal entropy increases monotonically as the 'quota' c rises (proof omitted due to space constraints), and it reaches the maximum,

$\log N$, when $c = c_m = \sum_{i=1}^{N}(p_1 - p_i)$. Afterwards, the optimal entropy stays at the maximum, since once the distribution has been made uniform further 'top-ups' can be made evenly to maintain uniformity. Similarly, the minimal entropy decreases as c rises monotonically and it approaches 0 when c goes to infinity. Typically $\mathcal{H}_T$ will be set to either $[(1-\theta_{\mathcal{H}})\mathcal{H}_0, (1+\theta_{\mathcal{H}})\mathcal{H}_0]$, the minimum needed to fall under the detector's radar. Note that this inverse problem solution can be used to calibrate the detector, since it provides the largest possible entropy 'response' corresponding to a distribution changing 'signal' of a given size.

Considering now the original problem, for $\mathcal{H}_T > \mathcal{H}$, the minimal value of the total increment required is unique because the inverse solution taking $[0, c_m] \mapsto [\mathcal{H}, \log N]$ is 1-1 onto. The increment should be spread over smallest probabilities to raise them to an uniform value. As for $\mathcal{H}_T < \mathcal{H}_0$, the minimal total increment is also unique because the inverse solution taking $[0, \infty) \mapsto (0, \mathcal{H}]$ is likewise 1-1 onto. The increment is entirely allocated to the largest probabilities. To actually solve for the minimal Δ, the entropy curve can be plotted as a function of Δ based on the solutions of **MaxH** and **MinH**. Then the optimal Δ can be quickly obtained by a numerical search.

Through solving the optimisation problems above, we obtain the technical results for camouflaging attacks from entropy based detection at minimal cost. These results also provide the insight into entropy's behaviour as a function of distribution shape.

4 Empirical Results

In this section we show how the results of the previous section can be used to answer core questions of interest to both the attacker and detector, such as whether an attack can be detected, and whether it can be disguised and at what cost. We begin with distributions from traffic traces, where we explore attacks on a single distribution and their camouflage, and then show how the camouflage technique can be extended to multiple distributions based on multiple traffic metrics. We then use idealised models to cleanly investigate a number of phenomena as a function of parameters. We focus on the case when the attack is *concentrated* on a single class index $i = t$, that is, $\mathcal{T} = \{t\}$. Nevertheless, the results below are generally valid for the concentrated attacks targeting a small number of indices.

4.1 Traffic Traces

We use 24 hours, from 00:00 to 23:59 March 30, 2009, of a 96-hour long trace captured from an OC-3 link, from the "Measurement and Analysis on the WIDE Internet" group (MAWI). The time series of interest were extracted using WireShark and our own C programs, with entropy calculations in MatLab. We focus mainly on a representative 5 minute time interval from 15:30 March 30, 2009. (5 minute intervals are commonly used, e.g. [7, 10]).

Concentrated attack detection

We take the packet count per destination IP address histogram (reordered),

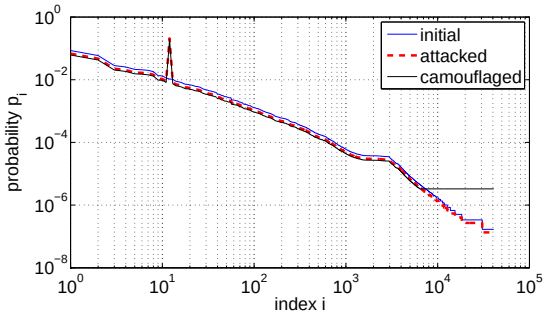

Fig. 2. Distributions of normal traffic, under the attack and after camouflage.

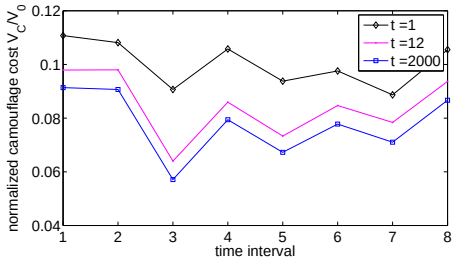

Fig. 3. Camouflage costs while attacking various targets with the intensity $V_A = 0.25V_0$ over 8 time bins with detection sensitivity $\theta_{\mathcal{H}} = 0.05$.

shown in Fig 2, as the benchmark distribution with $\mathcal{H}_0 = 8.92$ and $N = 40889$, and apply a concentrated attack with intensity $V_{\mathcal{A}} = 0.25V_0$ at the target index $t = 12$ (with $p_{12} = 0.0105$). This could arise for example as the result of a DoS attack when the packet metric is the destination IP, since then all V_A packets appear at a single index corresponding to the server under attack.

The attacker is aware that his attack has lowered the original entropy appreciably (by 12.48%). Seeking complete anonymity, he wishes to know the minimal number V_C of active camouflage packets needed to be invisible even to a perfect detector $\theta_{\mathcal{H}} = 0$. Since the attack has lowered entropy from $\mathcal{H}_0$ to $\mathcal{H}$, the camouflage packets must be placed so as to increase it back up to $\mathcal{H}_T = \mathcal{H}_0$. According to the camouflage scheme built in Section 3, the strategy reduces to the following: given the reordered version of $\mathcal{D}$, say $\mathcal{D}' = \{p'_i\}$, the opponent should increase the smallest $\bar{v}$ probabilities to the same value. In the case of Fig 2, which also shows the camouflaged solution $\mathcal{D}_C$ with $\mathcal{H}(\mathcal{D}_C) = 8.92$, $\bar{v} = 34527$ and $V_C = 0.135V$.

We also examined 8 intervals over the 24 hours with the same attack intensity, $V_A = 0.25V$, but various attacking targets, $t = [1, 12, 2000]$ with the sensitivity $\theta_{\mathcal{H}} = 0.05$. The results are similar to those for the representative interval above, and the costs over the 8 bins are shown in Fig 3. We observe that attacking any of these indexes results in an entropy drop. The larger the probability the greater the decrease, and so the higher the camouflage cost.

The above discussion is only an example. The same analysis is applicable to other concentrated attacks and detection based on other metrics. For example, a

worm attack may use fixed source port numbers, resulting in significant changes in a few indices of the source port distribution.

Complex detection scenarios

Some studies [7, 10] have considered the use of entropy of multiple distributions (e.g. destination IP addresses plus destination ports) in order to improve detection sensitivity. Specifically, an attack is declared if the entropy of any distribution under monitoring is out of its normal range. We now show that this does not increase the difficulty of camouflage compared to the single-metric case. We continue to use the 5 minute interval from 15:30 as our example.

Suppose that there is a DDoS attack with intensity $V_{\mathcal{A}} = 0.25V_0$ targeting index $t = 12$ of the (reordered) destination IP distribution and $t = 3$ of the (reordered) destination port distribution. Individually, the camouflage costs are $0.073V_0$ and $0.052V_0$ at sensitivity $\theta_{\mathcal{H}} = 0.05$. To evade detection based on the distribution pair, the camouflage cost is simply $0.073V_0 = \max\{0.073V_0, 0.052V_0\}$ because each packet can be used to camouflage either metric independently. The camouflage strategy for the address distribution is the same as that for the individual detection, whereas for the port distribution the attacker could use $0.052V_0$ camouflage packets to change the entropy to the desired value as before, and then use the remaining $0.021V_0$ packets to improve the port-camouflage further. In a similar way, the camouflage technique can be applied to other scenarios with more complex detection mechanisms such as in [5].

4.2 Synthetic Distributions

We now provide a more systematic investigation of the attacker-detector battle using a simplified model distribution, specifically a truncated Zipf with $s = 1.5$, $N = 10^4$ and $\mathcal{H}_0 = 4.47$.

Imperfect detector Clearly, with less sensitive detectors the camouflage possibilities grow, as seen in Fig 4(a), which gives an example of how, for a fixed target $t = 12$ and for each of several different sensitivity levels, the camouflage cost $V_{\mathcal{C}}$ varies as a function of the attack intensity. Not surprisingly, the range of intensities where the cost is zero increases monotonically with $\theta_{\mathcal{H}}$. When $\theta_{\mathcal{H}} = 0$ this is only possible at a single value of intensity ($V_{\mathcal{A}} = 0.09V_0$), but the range expands to $(0, 0.17]$, $(0, 0.25]$ as $\theta_{\mathcal{H}}$ rises through $0.02, 0.05$ respectively. For a fixed attack intensity, whenever camouflage is needed, the volume of camouflage required is monotonically decreasing in $\theta_{\mathcal{H}}$.

Intuitively, we expect that concentrated attacks lower entropy since they concentrate probabilities, but this is not always true. The attack may cause an entropy rise when it is moderate. The active camouflage cost for $V_{\mathcal{A}} \in (0, 0.09V_0)$ and $\theta_{\mathcal{H}} = 0$ in Fig 4(a) is an example. In addition, camouflage volume is monotonic in attack intensity when the resulting entropy reduces. But this monotonicity does not hold when the entropy increases.

Relative costs Sometimes the attacker may be more interested in the marginal cost of camouflage rather than the absolute. Fig 4(b) gives the relative cost corresponding to Fig 4(a). We see that for highly intensive attacks like $V_{\mathcal{A}} = V_0$ the relative and absolute costs tell a similar story. However, when the attack

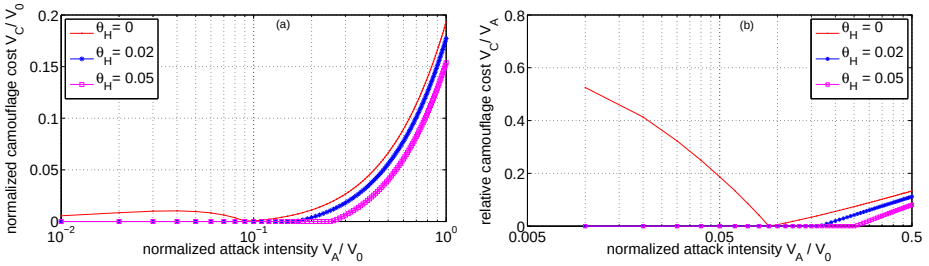

Fig. 4. Camouflage costs when attacking $t = 12$ for different detector sensitivities. (a): absolute cost. (b): relative cost.

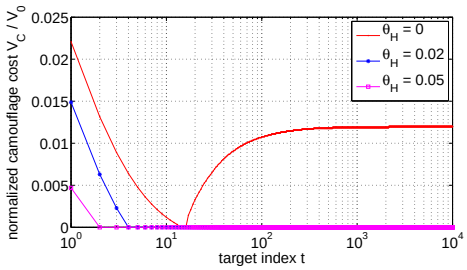

Fig. 5. Camouflage costs while targeting different indices with $V_{\mathcal{A}} = 0.1V_0$.

size is small, trying to protect oneself against a perfect detector becomes very expensive relative to the attack volume.

Various targets Finally, we study the camouflage cost as a function of the target t at a constant attack intensity. Fig 5 shows the costs for $V_{\mathcal{A}} = 0.1V_0$ targeting different indices. Entropy decreases as the result of an attack for small t's, and then the behaviour of $V_{\mathcal{C}}$ is simple and monotonic in both t and $\theta_{\mathcal{H}}$. Once t is large enough, the entropy rises rather than drops, and then the camouflage strategy is no longer to raise probabilities in the tail, but to increase the largest: p_1. However, a given absolute change in p_1 influences entropy less than the same change at smaller probabilities, resulting in a larger value of $V_{\mathcal{C}}$ being required to rebalance the entropy. Consequently, the camouflage cost goes up significantly, in particular for a sensitive detector.

In our examples we assumed the attacker was capable of mounting an attack with relatively large $V_{\mathcal{A}}$, which could constitute a very large amount of traffic on high capacity links. Even then we saw that an $\mathcal{H}$ based detector often failed to detect these attacks. If $V_{\mathcal{A}}$ is much smaller, which is more realistic in many contexts, the attack will be much harder to detect. In any event, we provide the framework and technical results needed to explore these and other related issues.

5 Conclusions and Future Work

We have examined the behaviour of Shannon entropy as a summary statistic, and pointed out that it suffers from a number of significant weaknesses in the context of network attack detection. We formulated and solved optimisation problems

yielding the first rigorous results on 'optimal camouflage'. These are of relevance both to detectors and attackers to understand how entropy signatures can be either passively or actively reduced, and to evaluate the cost required to make them invisible to detectors.

Attack and detection strategies are subject to an arms race. We have provided the underlying tools essential to analyse both sides of the battle in simple scenarios, and have shown how more complex cases can be built up using them. We hope our generic approach will be useful as a foundation for the development of new detectors against ever more sophisticated attackers.

There are many directions for future work. These include allowing both decrement and increment based camouflage and discussing distributed attacks. Other questions of interest include investigating optimal camouflage strategies when the attacker has only limited information about the benchmark distribution underlying the detection, and how to overcome the limitations of entropy.

References

1. Celenk, M., Conley, T., Willis, J., Graham, J.: Anomaly Detection and Visualization using Fisher Discriminant Clustering of Network Entropy. In: Pichappan, P., Abraham, A. (eds.) Third IEEE ICDIM, pp. 216–220 (2008)
2. Dalvi, N., Domingos, P., Mausam, Sanghai, S., Verma, D.: Adversarial Classification. In: KDD '04: Proceedings of the 10th ACM SIGKDD, pp. 99–108 (2004)
3. Feinstein, L., Schnackenberg, D., Balupari, R., Kindred, D.: Statistical Approaches to DDoS Attack Detection and Response. In: DARPA Information Survivability Conference and Exposition, vol. 1, pp. 303–314 (2003)
4. Lee, W., Xiang, D.: Information-Theoretic Measures for Anomaly Detection. In: Proc. of IEEE Symposium on Security and Privacy, pp. 130–143 (2001)
5. Li, L., Zhou, J., Xiao, N.: Information and Communications Security, S. Qing, H. Imai, G. Wang, editors, chap. DDos Attack Detection Algorithms based on Entropy Computing, pp. 452–466. Springer - Verlag Berlin Herdelberg (2007)
6. Lowd, D., Meek, C.: Adversarial learning. In: KDD '05: Proceedings of the 11th ACM SIGKDD, pp. 641–647 (2005)
7. Nychis, G., Sekar, V., Andersen, D.G., Kim, H., Zhang, H.: An Empirical Evaluation of Entropy-Based Traffic Anomaly Detection. In: IMC '08: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement, pp. 151–156 (2008)
8. Rubinstein, B.I., Nelson, B., Huang, L., Joseph, A.D., Lau, S.h., Rao, S., Taft, N., Tygar, J.D.: ANTIDOTE: Understanding and Defending Against Poisoning of Anomaly Detectors. In: IMC '09: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, pp. 1–14 (2009)
9. Tellenbach, B., Burkhart, M., Sornette, D., Maillart, T.: Beyond Shannon: Characterizing Internet Traffic with Generalized Entropy Metrics, pp. 239–248. Springer - Verlag Berlin Herdelberg (2009)
10. Wagner, A., Plattner, B.: Entropy based Worm and Anomaly Detection in Fast IP Networks. In: WETICE '05: Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, pp. 172–177 (2005)
11. Ziviani, A., Gomes, A.T.A., Monsores, M.L., Rodrigues, P.S.S.: Network Anomaly Detection using Nonextensive Entropy. IEEE Communications Letters **11**(12), 1034–1036 (2007)