A New, Principled Approach to Anomaly Detection

Erik M. Ferragut, Jason Laska, Robert A. Bridges
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, USA
{ferragutem, laskaja, bridgesra}@ornl.gov

***Abstract*—Intrusion detection is often described as having two main approaches: signature-based and anomaly-based. We argue that only unsupervised methods are suitable for detecting anomalies. However, there has been a tendency in the literature to conflate the notion of an anomaly with the notion of a malicious event. As a result, the methods used to discover anomalies have typically been ad hoc, making it nearly impossible to systematically compare between models or regulate the number of alerts. We propose a new, principled approach to anomaly detection that addresses the main shortcomings of ad hoc approaches. We provide both theoretical and cyber-specific examples to demonstrate the benefits of our more principled approach.**

I. Introduction

Cyber security analysts are confronted with enormous data sets produced from a myriad of sources. From this complex of low level logs and high level alerts, they are charged with discovering malicious behavior, behavior which can come from various sources and make use of a diversity of methods. To help address this overwhelming challenge, cyber analysts are keen to make use of any automated tools available.

Tools for attack detection are traditionally viewed as being either signature-based or anomaly-based. Signature-based approaches look for particular cues that have been linked, through past observations, to previously identified attacks. Anomaly-based approaches attempt to identify events that are unusual and, presumably, more likely to be malicious.

Anomaly detection has, therefore, two demands placed upon it. First, it should detect the unusual events. Second, these unusual events should correlate with malicious events. A careful reading of the literature shows that while the first demand motivates the approaches taken, it is the second demand that is used to judge the efficacy of anomaly detection [2].

We argue that these two demands should not both be addressed by the same approach. Attempting to do so inevitably conflates anomalousness with maliciousness, which results in a weaker anomaly detector and also becomes inherently a signature-based detection scheme. The literature generally has not viewed anomaly-malice conflation as a problem. This is evidenced even in Denning [4], where cyber security anomaly detection was first proposed: "An important objective ... is to determine what activities and statistical measures ... have a high rate of detection and a low rate of false alarms."

The primary philosophical problem with training an anomaly detector to have good discriminating power for detecting malicious attacks is the unavoidable use of supervised learning. As is well known in the machine learning community, a trained classifier will generalize to other examples similar to those in the training set, but not to entirely new types of examples. As such, a so-called anomaly detection method that is trained to detect maliciousness is necessarily a signature-based approach, albeit one with a potentially broad and robust definition of signature. While this may contrast sharply with the traditional use of signature detection, which typically refers to scanning for particular byte sequences, it is nevertheless a trained model for detecting a particular class of behavior.

We propose that the more precise terminology of machine learning be brought to bear on the definitions of signature- and anomaly-based methods. In particular, signature-based methods are supervised learning methods: they develop means of detecting malicious behavior on the basis of *previously seen* malicious events. Anomaly-based methods are (or rather, should be) unsupervised methods: they develop means of detecting *atypical* events. This distinction has profound implications for cyber security. After all, high quality (i.e., accurate, representative, large-scale) labeled data sets in cyber security are exceedingly rare outside of malware and packet captures of known attacks. The practice of signature-based approaches can be used to build models to detect these known attacks (either by direct observation or by detection of the side effects in computer behavior). This is a very important aspect of cyber security. However, the threat of zero-day attacks looms large in the cyber security landscape. Detecting new and unexpected events cannot properly be done using a classifier as it is impossible to train a class on unseen examples. Instead, unsupervised methods (including one-class classifiers) are needed. In the cyber security domain, unsupervised methods have the key advantages of not requiring labeled data and allowing for zero-day attack detection.

Methods of unsupervised learning, such as density estimation [1], clustering [5], [7], and dimension reduction [6], have been applied to cyber security. Two common challenges arise in practice. First, methods are needed to deal with disparate sources of data, such as network flow data, firewall logs, and system logs. A method is needed to provide *comparability* of, say, a network flow anomaly to a firewall log anomaly, which

The manuscript has been authored by a contractor of the U.S. Government under contract DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

is challenging if they adopt different or ad hoc anomaly detection approaches. Second, the methods must be *regulatable* in the sense that analysts can in advance set the proportion of false alerts. (A definition for false alerts in the context of unsupervised learning is in order. The anomaly detection method will be applied to two types of data. The first is data produced similarly to the training data. The second is data produced by a different, unknown method. Data produced similarly to the training data but that is flagged as anomalous is a false alert. The rate of false alerts does not depend on the choice of the second source of data since the model was trained without access to that source.) This is especially important in cyber security because of the size of the data sets. A false alert rate of 0.001 may seem small, but in a network with one million events per hour, there would be 1000 false alerts per hour. Of course, reducing the false alert rate inevitably reduces the true alert rate as well. This problem is addressed by a good choice of model, but is outside the scope of this paper.

In the remainder of this paper we introduce a new, principled approach to anomaly detection in the context of existing methods, and provide both theoretical and practical examples. In Section II, we consider existing notions of anomalousness and highlight some of their shortcomings. In Section III, we provide our own, new definition of anomalousness and state our main theorem. Section IV provides some theoretical examples that show how our definition agrees with commonsense notions of anomalousness and addresses the shortcomings of existing methods. Section V provides some practical applications of how our definition has been used in cyber security. We conclude in Section VI.

II. Issues with Existing Anomaly Definitions

Informal notions of anomalousness are commonly used in cyber security and in other domains. We consider three interpretations of what constitutes an anomaly. First, an event may be viewed as anomalous if it is *abnormal* (i.e., contrary to normative expectations). As stated above, this perspective leads naturally to a supervised learning problem since normality must be described. Since supervised learning lacks the benefits of anomaly detection in cyber security, we do not pursue this perspective further. Second, an event is anomalous if it is *rare* (equivalently, atypical, unusual). This perspective requires a means of determining how common an event is. Third, an event is anomalous if it is *different* (equivalently, peculiar, strange). This perspective requires a means of determining whether an event is similar to others. In this section, we discuss methods that pursue the second and third of these three interpretations. We argue that they are both answered using a probability distribution, whether it be explicit, as in the case of measuring rarity, or implicit, as in the case of measuring difference. As we will make probability distribution assumptions explicit, we will refer to this unified concept of anomalousness as rarity. In Section III, we demonstrate that simply using the overall rarity provides neither the regulatability nor the comparability that we require of a useful anomalousness score.

A. Anomalies as Rare Events

As a typical, recent example of a rarity perspective of anomalies, we review Tandon and Chan [8]. They model users' geographic locations over times using Markov chains (of orders zero and one) using discrete times and locations. This provides explicit probabilistic models of the users' positions over time. For each probability distribution, they construct an anomaly score of a location in a given temporal window by taking the negative log of the probability. They define an anomaly to be an event having an anomaly score that exceeds a threshold. They note that this model will "flag valid but low frequency events as anomalous, resulting in higher false alarms." Consequently, they develop two more advanced and better performing probabilistic models. This is an example of the common conflation of anomalousness and maliciousness, where a model trained to detect *anomalous* events is validated by its ability to detect *malicious* events. As a result, the definition was modified to better capture their pre-conceived notion of maliciousness. Tandon and Chan's definition of anomalousness is sufficiently common in the literature that we will give it its own definition.

Definition 2.1 (Bits of Rarity): Given a probability distribution described by probability density or probability mass function f, the anomaly score of an event x is given by

$$\mathrm{R}_f(x) = -\log_2(\mathrm{P}_f(x)).$$

Using the log helps with numerical stability of the computation; the negative ensures that the most atypical events have the highest anomaly scores.

Unfortunately, in general, the bits of rarity definition is neither regulatable nor comparable. We demonstrate these shortcomings using theoretical examples in the next section. Tandon and Chan [8] construct a ROC curve and select a threshold with a given, preselected false alarm rate. This is analogous to the practical approach of ranking anomalies and establishing a threshold that selects a given proportion of events as anomalous. This approach is feasible for post hoc analysis, but cannot realistically be applied to streaming data, which is the common operational case in cyber security. Tandon and Chan do not need to deal with comparability since they only consider anomalies from one type of data.

B. Anomalies as Different Events

Portnoy et al. [7] provide an analysis typical of the perspective of anomalies as *different*. They construct vectors from network traffic where each coordinate (i.e., feature) is independently normalized to have mean zero and standard deviation one. After clustering the vectors, the smallest clusters are considered anomalous. Effectively, an event is viewed as being *different* if it is not assigned to any of the largest clusters. This approach also lacks regulatability and comparability. Similarly to Tandon and Chan [8], Portnoy et al. determine their threshold by post hoc analysis.

One main difference between the *rarity* and *difference* perspectives of anomalousness is that *rarity* requires an explicit notion of frequency or probability whereas *difference* does not. However, since *difference* is computed with respect to typical events, an *implicit* notion of probability is required. In the clustering approach of Portnoy et al., we can consider a mixture of Gaussians with one component for each cluster. For each event, x, we define c_x to be the closest cluster. For a random event, X, its cluster assignment C_X is a random multinomial variable. (The parameters of the multinomial would be approximately the mixture probabilities of the Gaussian mixture model.) The anomaly score they used considers the rarest of these clusters to be the anomalies. Therefore, it is approximately the same as applying the bits of rarity definition to C_x.

The key difference between *rarity* and *difference* perspectives of anomalousness is therefore whether the probabilistic model is given explicitly or implicitly. It is our view that making one's assumptions explicit is the superior choice when possible since it allows the assumptions to be tested and revised. Furthermore, explicit probabilistic distributions enable the derivation of Bayesian optimal inference rules, which are likely to outperform approximate heuristic approaches.

Unfortunately, the bits of rarity definition of anomalousness is generally neither regulatable nor comparable. In the next section, we propose a principled, probabilistically motivated improvement that addresses these issues.

III. An Improved Definition of Anomalousness

In this section, we propose a definition of anomalousness based on the probability *of the probability*. In the previous section, we defined bits of rarity as a measure of how unlikely an event is, noting that it is a commonly used approach in the literature. Unfortunately, bits of rarity is not sufficient, as it lacks both regulatability and comparability. This can be demonstrated by considering the threshold selection problem. If a threshold of, say, 10 is set in advance, then an event with a probability 2^{-10} or lower would be considered anomalous. Now consider a uniform discrete distribution. If it has 100 possible values, then none of the events are considered anomalous. However, if it has 2000 possible values, then they are each considered anomalous *even though they are each maximally likely*. A regulatable definition of anomalousness would be self-adapting to these various distributions. Intuitively, it is not just the rarity of the event itself, but how surprising that rarity is in the context of the distribution.

Similar to regulatability, the bits of rarity definition lacks comparability. In practical applications, data come from a multiplicity of sources, each with their own properties. Suppose that we observe two types of discrete variables, one has two possible values and another that has a thousand possible values. An ideal definition of anomalousness would apply to both distributions without requiring the tuning of a parameter for each dataset, and yet would allow for the direct comparison of the anomalousness of values of one variable with values of the other variable. The bits of rarity definition does not allow for this, but direct comparison can be accomplished by considering not the rarity of the event itself, but how *rare the rarity of an event is*.

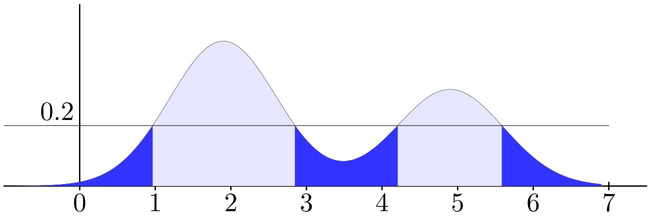

Fig. 1. The anomalousness of an event x is the negative log base two of the area under the PDF curve restricted to those t such that $f(t) \leq f(x)$.

Definition 3.1 (Bits of Meta-Rarity): Given a (discrete or continuous) random variable X with probability density or mass function f defined on domain $\mathcal{D}$, let $A_f : \mathcal{D} \to \mathbb{R}_{\geq 0}$

$$A_f(x) := -\log_2 \mathrm{P}_f(f(X) \leq f(x))$$

be the *anomalousness* or *anomaly score* of x.

Remarks. Note that A_f is defined on $\mathcal{D}$, the same domain as f. Again, we have adopted the negative log for numerical reasons and so that larger anomalousness corresponds to larger numbers. Also, since the probabilities of interest are likely to be very close to zero, the use of log helps emphasize their differences. The choice of base 2 for the log is chosen so that the anomalousness is, in an information theoretic sense, measured in bits. By convention, we define the negative log of zero to be positive infinity. The subscript f in A_f may be omitted if clear from context.

Our new definition can be interpreted with respect to the graph of the probability density, as shown in Figure 1. Given an event x, $\mathrm{P}_f(f(X) \leq f(x)) = \int_{\{t|f(t) \leq f(x)\}} f(t)\mathrm{d}t$, hence $\mathrm{P}_f(f(X) \leq f(x))$ equals the area of the shaded region in Figure 1. The negative log base two of that area is the bits of meta-rarity (i.e., anomaly score).

We now present a brief example of the definition; additional examples will be provided in the next two sections. Consider the discrete uniform distribution. In this case, each event has the same probability. Hence, for any x, the probability of the random variable X having probability mass $f(X)$ less than or equal to $f(x)$ is one. Therefore, $A_f(x) = -\log_2 1 = 0$ for all x in the distribution. (This is a particular case of the general observation that any mode of a distribution has anomaly score 0.) Without selecting a threshold or tuning a parameter, we can conclude that a discrete uniform distribution has no anomalies. This agrees with intuition, since an event that is just as likely as every other event should be considered typical.

A significant advantage of this definition of anomaly is that $A_f(X)$ for a random variable X is predictable, and $P_f(A_f(X))$ can be bounded independent of f under the assumption that X is generated according to the distribution described by f. We prove this in the following theorem. Regulatability and comparability of A_f then follow.

Theorem 3.2: Let X be distributed according to probability distribution f. Then the probability that the anomalousness exceeds α is no greater than $2^{-\alpha}$. That is,

$$\mathrm{P}_f(A_f(x) \geq \alpha) \leq 2^{-\alpha}$$

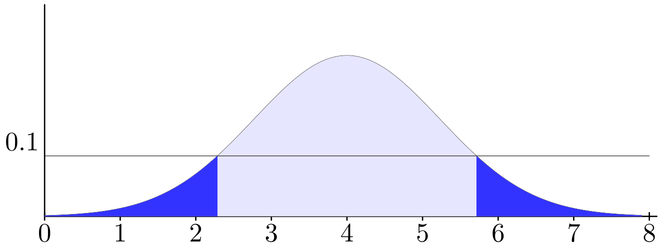

Fig. 2. The bits of meta-rarity definition of anomalousness is monotonically equivalent to the z-score.

The proof of the theorem is given in the Appendix.

By the theorem, the proportion of events flagged as false alerts at the α level will be no more than $2^{-\alpha}$ for samples generated according to f. In particular, the number of false alerts at a given threshold is independent of f. Hence, *false alerts can be regulated* by selecting an appropriate α. Furthermore, if X is produced according to f and Y is produced according to g, then $A_f(X)$ and $A_g(Y)$ are comparable since they are both negative log probabilities. *This definition of anomalousness therefore provides comparability across different sources* even if each source is modeled using a different probability distribution.

Note that the bits of meta-rarity definition of anomalousness has no parameters that need to be set arbitarily or by experimentation. The definition is, in this sense, self-tuning: it uses the distribution itself as sufficient context for determining anomalousness. One reasonable way to use these advanatages of our definition is to set a threshold based on the size (or throughput) of the data to be analyzed. If a cyber security data set has, say, one million events that will be scored for anomalousness, then setting a threshold at $\log_2 1,000,000 = 19.93 \approx 20 = \alpha$ should yield at most one false alert assuming that events are really generated according to f by the theorem. Deviations in the number of anomalies will indicate that the model (i.e., choice of f) does not match the generating distribution. This mismatch could be because f was not properly selected or tuned, or it could be because there is another source of events. In either case, exploration of the anomalies will provide insight into both the state of the system and changes within it.

IV. Theoretical Examples

Example 4.1 (Gaussian): First we consider the Gaussian distribution, where our definition of anomalousness is (monotonically) equivalent to a z-score. The z-score essentially captures the normalized distance to the mean and offers regulatability since the distributions are known. However, the z-score is specific to Gaussian distributions, making comparability across different distributions difficult. Figure 2 shows how our definition picks out the tails of a Gaussian distribution, in agreement with a z-score-based definition of anomalousness.

We now work out this example in detail. For a Gaussian distribution with mean μ and standard deviation σ, the probability density function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is defined as

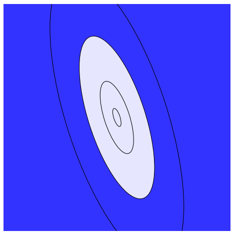

Fig. 3. The level curves of the probability density function of a multivariate Gaussian are concentric ellipses. Our definition of anomalousness generalizes to multivariate distributions, and is monotonically equivalent to Mahalanobis distance for multivariate Gaussians.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The anomalousness is then given by

$$A_f(x) = -\log_2 P_f(f(X) \leq f(x)).$$

We see that $f(y) \leq f(x)$ if and only if $(y-\mu)^2 \geq (x-\mu)^2$. The probability of the set of such y is then the sum of the tail probabilities, which can be given in terms of the cumulative distribution function F. The left tail has probability $F(\mu - |x-\mu|)$ and the right tail has probability $1 - F(\mu + |x-\mu|)$. However, by the symmetry of the Gaussian distribution, these two probabilities are equal. Hence the anomalousness can be written as

$$A_f(x) = -\log_2(2F(\mu - |x-\mu|)).$$

Typically, the observations are first standardized by defining $z = \frac{x-\mu}{\sigma}$. Let G denote the cumulative distribution function of the standardized one-dimensional Gaussian. Then the anomalousness becomes

$$A_f(z) = -\log_2(2G(-|z|)).$$

Evidently, the more x deviates from μ the more anomalous. Our anomaly score is therefore monotonically equivalent to the z-score. However, an important distinction is that anomalousness A_f does not require the parameteric assumption of the z-score.

For a given false alert rate and a given Gaussian distribution, an appropriate threshold can be deduced. This shows that the bits of rarity definition provides regulatability in this case. However, the threshold depends explicitly on the distribution parameters. Bits of rarity (at a given threshold) gives different false alert rates for different parameters. Hence the bits of rarity definition does not provide comparability across distributions.

Example 4.2 (Multivariate Gaussian): For a k-dimensional multivariate Gaussian, the probability distribution function $f : \mathbb{R}^k \to \mathbb{R}$ is defined to be

$$f(x) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)\right)$$

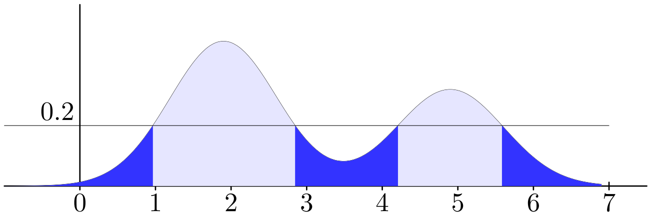

Fig. 4. Although the z-score works well for a Gaussian distribution, it does not generalize to other distributions. Our definition captures the rarest events, which is a generalization of the notion of tails.

where μ is the mean and Σ is the positive definite covariance matrix. (Here, x and μ are thought of as column vectors, and v^t for v a column vector is the transpose of v, which is a row vector.)

Note that f is monotonic in $-(x-\mu)^t\Sigma^{-1}(x-\mu)$. The level sets of the distribution are depicted in Figure 3. The anomalousness of an event x is then the probability of an event being outside that level curve. (Note that this observation uses the unimodality of f.) Since Σ is positive definite, its square root $S = \sqrt{\Sigma}$ can be computed such that $\Sigma = S^t S$. Then,

$$A_f(x) = -\log_2 P_f\left(\|S^{-1}(X-\mu)\| \geq \|S^{-1}(x-\mu)\|\right).$$

We have shown again that our definition of anomalousness agrees with common practice, as it identifies the tails. In fact, this shows that the anomalousness of a multivariate Gaussian event is monotonically equivalent to its Mahalanobis distance, a common reparameterization used in machine learning. Mahalanobis distance, like the z-score, can be used to provide regulatability, but fails to provide comparability across distributions.

Example 4.3 (Gaussian Mixture): The two previous examples show that the distance to the mean (appropriately normalized) provides a reasonable definition of anomalousness for some distributions. However, it is problematic for multimodal distributions. The Mixture of Gaussian distributions has a probability density function that is the weighted sum of multiple Gaussian distributions, as shown in Figure 4. Potentially, the mean of a Mixture of Gaussian distribution could be an anomaly, since it can fall in a valley between the modes, as in the figure. This example illustrates that a general definition of anomalousness cannot be based on identifying just the tails. The bits of rarity and the bits of meta-rarity definitions both capture the rare events in the middle of the distribution as anomalous.

Example 4.4 (Multinomials): Bits of meta-rarity applies equally well to discrete distributions. For these, the anomalousness of x is the log base two of the sum of all probabilities less than or equal to $\mathrm{P}_f(x)$. This is demonstrated in Figure 5. Because of the comparability, we can meaningfully compare the anomalousness of a multinomial variable with the anomalousness of, say, a Gaussian mixture variable.

One advantage of our approach is that it extends to any random variable, inclusive of complex probabilistic constructions, such as random graphs and stochastic processes.

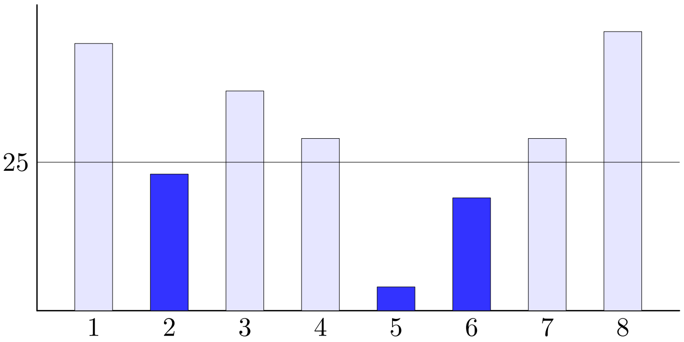

Fig. 5. Our new definition of anomalousness applies equally well to continuous and discrete random variables

	Destination				
Source	W	F	D	S	U
W	0	0.3667	0	0.6254	0
F	0.0000003	0	0	0	.00008
D	0.000012	0	0.0000317	0	0.0000461
S	0	0.0013673	0	0	0.0003282
U	0	0.0019842	0	0.00386	0.0001351

TABLE I
PROBABILITIES FOR COMMUNICATIONS BETWEEN IP ROLES.

V. CYBER SECURITY EXAMPLES

In this section we apply our approach specifically to a cyber security data set. For this, we selected the VAST 2012 Mini-Challenge 2 data set [3]. The data set provides firewall logs wherein each entry is comprised of a timestamp, source IP address, source port, destination IP address, destination port, protocol, and message code. First, we define a variable derived from a log. Second, we estimate its probability distribution. Third, we score events for anomalousness.

Example 5.1 (IP-to-IP traffic by role): For a given observation we first extract the pair

$$(\text{source IP role}, \text{destination IP role}).$$

These roles were taken from the information provided with the challenge scenario. In particular, for each IP address, we knew its assigned role. The possible roles were Workstation, Firewall, DNS, Web Server, and Unassigned. (We abbreviate these as W, F, D, S, and U.)

We take any such pair as the observed value of a random multinomial variable X. Let $N_{a,b}$ be the number of firewall log entries with source IP role a and destination IP role b. We estimate the probabilities $f((a,b))$ by $N_{a,b}/\sum_{x,y} N_{x,y}$. (Incorporating priors would assist with the scoring of previously unseen events.) The anomalousness of a pair (a,b) is given by

$$A_f((a,b)) = -\log_2 \sum_{\{(x,y)\,|\,f(x,y)\leq f(a,b)\}} f(x,y).$$

The observed probabilities and anomaly scores for the random variable are summarized in Tables I and II.

The most anomalous communication originates from the firewall and terminates at a workstation with an anomalous

	Destination				
Source	W	F	D	S	U
W		1.417		0	
F	21.669				12.525
D	16.311		14.472		13.438
S		8.966			10.626
U		7.971		6.994	11.68

TABLE II
ANOMALY SCORES FOR COMMUNICATIONS BETWEEN IP ROLES.

score of 21.669. Indeed traffic classified in this group represents communication specifically between the firewall and the log server. On one hand, the relative lack of communication conforms with our expectations of standard network behavior; however, further analysis indicates that communication between the firewall and the log server surprisingly terminates after 15 minutes of the 40 hour dataset.

The second most anomalous communication originates from the DNS server and terminates at a workstation with an anomalous score of 16.311. Further analysis of this traffic indicates attacks on the DNS server involving ports 135, 137, 139, and 445, all of which are associated with Microsoft file-sharing traffic.

The third most anomalous communication originates from the DNS server and terminates at a DNS sever with an anomalous score of 14.472. Further analysis of this traffic indicates normal DNS traffic. This is expected network behavior. It is noteworthy that the traffic from DNS to DNS is more anamalous than traffic from DNS to Unlabeled. This trend indicates a possible loss of control of the DNS server.

In this example, our anomaly scoring served to identify atypical events in a streaming environment. The insights by this process provided a useful first step in developing a comprehensive situational understanding of the network.

In application, the main challenges of our new definition are those that typically arise when working with probability distributions. Determining the proper class of distributions to use, tuning the parameters of the distribution, dealing with changes in the distribution over time, and developing computationally efficient algorithms can all be challenges dependening on the application. On the other hand, methods that do not explicitly define a probability distribution are making implicit assumptions that are impossible to validate.

VI. CONCLUSION

We have defined a principled probability-based definition of anomalousness that is reasonable (in that it agrees with intuition in typical examples), general (in that it applies to anything modeled by a probability distribution), comparable (in that scores of disparate types of events can be compared), and regulatable (in that the rate of false alerts can be set in advance). Together, these benefits demonstrate that our definition addresses the main shortcomings of previous methods. We employed our anomalous score on both theoretical and applied examples to show that it agrees with intuition, generalizes to many cases, and conforms to real-world requirements.

APPENDIX

Adopting measure theory notation

$$A_f(x) = -\log_2(\mu\{t : f(t) \le f(x)\}).$$

Note that

$$\begin{aligned}&\{x : \mu\{t : f(t) \le f(x)\} \le \mu\{t : f(t) \le f(y)\}\} \\ &= \{x : f(x) \le f(y)\}\,.\end{aligned}$$

Proposition A.1: Fix $y \in \mathcal{D}$. Then

$$\mathrm{P}(A_f(X) \ge A_f(y)) = 2^{-A_f(y)}.$$

Proof: $\mathrm{P}(A_f(X) \ge A_f(y))$ may be rewritten as

$$\begin{align} &= \mu\{x : A_f(x) \ge A_f(y)\} \notag \\ &= \mu\{x : -\log_2(\mu\{t : f(t) \le f(x)\}) \notag \\ &\qquad\qquad \ge -\log_2(\mu\{t : f(t) \le f(y)\})\} \notag \\ &= \mu\{x : \mu\{t : f(t) \le f(x)\} \le \mu\{t : f(t) \le f(y)\}\} \tag{1} \\ &= \mu\{x : f(x) \le f(y)\} \notag \\ &= 2^{-A_f(y)}, \notag \end{align}$$

which proves the proposition. ■

Proof of Main Theorem: We recall that $\mathrm{P}(A_f(X) \ge \alpha) = \mu\{x : A_f(x) \ge \alpha\}$. We break the proof into two cases. Case 1: Suppose that for all $y \in \mathcal{D}, A_f(y) < \alpha$. Then we see the result is trivially true as $\{x : A_f(x) \ge \alpha\} = \emptyset$. Case 2: Now we suppose that $\{x : A_f(x) \ge \alpha\} \ne \emptyset$, so that there exists some y such that $A(y) \ge \alpha$. Then we set $r = \inf\{A_f(x) : A_f(x) \ge \alpha\}$, and let $x_n \in \mathcal{D}$ so that $A_f(x_n) \downarrow r$. Hence,

$$\{x : A_f(x) \ge \alpha\} = \bigcap_{n=1}^{\infty} \{x : A_f(x) \ge A_f(x_n)\},$$

the sets on the right being nested. By the finiteness of the measure,

$$\begin{align} \mu\{x : A_f(x) \ge \alpha\} &= \lim_n \mu\{x : A_f(x) \ge A_f(x_n)\} \notag \\ &= \lim_n 2^{-A_f(x_n)} \tag{2} \\ &= \lim_n 2^{-r} \notag \\ &\le 2^{-\alpha}, \notag \end{align}$$

since $r \ge \alpha$, where (2) follows from Proposition A.1. ■

REFERENCES

[1] Y. Cao, H. He, H. Man, and X. Shen. Integration of self-organizing map (som) and kernel density estimation (kde) for network intrusion detection. In *Proc. of SPIE Vol*, volume 7480, pages 74800N–1, 2009.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[3] K. Cook, G. Grinstein, and M. Whiting. Vast challenge 2012. http://www.vacommunity.org/VAST+Challenge+2012, Aug. 2012.

[4] D. Denning. An intrusion-detection model. *Software Engineering, IEEE Transactions on*, 2(2):222–232, 1987.

[5] G. Gu, R. Perdisci, J. Zhang, and W. Lee. Botminer: Clustering analysis of network traffic for protocol-and structure-independent botnet detection. In *Proceedings of the 17th conference on Security symposium*, pages 139–154. USENIX Association, 2008.

[6] L. Huang, X. Nguyen, M. Garofalakis, M. Jordan, A. Joseph, and N. Taft. In-network pca and anomaly detection. *Advances in Neural Information Processing Systems*, 19:617, 2007.

[7] L. Portnoy, E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001*. Citeseer, 2001.

[8] G. Tandon and P. K. Chan. Tracking user mobility to detect suspicious behavior. In *SDM*, pages 871–883, 2009.