Challenging the Anomaly Detection Paradigm A provocative discussion

Carrie Gates
CA Labs
Islandia, NY
carrie.gates@ca.com

Carol Taylor
Faculty of Computer Science
University of Idaho
Moscow, Idaho
ctaylor@cs.uidaho.edu

ABSTRACT

In 1987, Dorothy Denning published the seminal paper on anomaly detection as applied to intrusion detection on a single system. Her paper sparked a new paradigm in intrusion detection research with the notion that malicious behavior could be distinguished from normal system use. Since that time, a great deal of anomaly detection research based on Denning's original premise has occurred. However, Denning's assumptions about anomalies that originate on a single host have been applied essentially unaltered to networks. In this paper we question the application of Denning's work to network based anomaly detection, along with other assumptions commonly made in network-based detection research. We examine the assumptions underlying selected studies of network anomaly detection and discuss these assumptions in the context of the results from studies of network traffic patterns. The purpose of questioning the old paradigm of anomaly detection as a strategy for network intrusion detection is to reconfirm the paradigm as sound or begin the process of replacing it with a new paradigm in light of changes in the operating environment.

Keywords

security, intrusion detection

1. INTRODUCTION

Intrusion detection is an important defence mechanism used by defenders to determine if someone has penetrated their system. Two approaches have typically been taken when designing intrusion detection systems: signature-based and anomaly detection. Signature-based systems, such as Snort [25], match incoming packets against various signatures that represent different types of malicious activity, such as particular buffer overflow attacks or signatures for worms. Unfortunately, such a system is reactive in that a malicious activity must first exist before a signature can be developed. Anomaly detection attempts to address this short-coming by alerting on changes in activity, where these changes are unusual (anomalous).

A great deal of research effort has gone into creating anomaly detection systems, although very few systems have seen widespread use. Such systems have been developed to operate at the host level to detect if a user is attempting to abuse an application in order to gain root privileges (*e.g.*, Forrest *et al.* [11]), and at the network level to detect if a remote adversary is attempting to gain unauthorized access (*e.g.*, MInDS [10]). However, little work has gone into determining if the underlying assumptions hold. In particular, it is assumed that malicious behaviour is anomalous, and therefore that by detecting anomalous behaviour we are detecting malicious behaviour. This assumption was first introduced by Dorothy Denning in her landmark paper on the subject. In this paper [8], she states "... exploitation of a system's vulnerabilities involves abnormal use, of the system; therefore, security violations could be detected from abnormal patterns of system usage."

While this assumption was perhaps correct in 1987, when the main concern was detecting intrusions within a single system, we believe the assumption of anomalous-equals-malicious breaks down in the highly variable, networked systems of today. In fact, we question whether anomaly detection is even a viable approach for a network environment given the well documented difficulties of characterizing Internet traffic [24]. Yet the appeal of anomaly detection along with the unquestioning acceptance of Denning's work continues to influence research within the security community despite significant differences in the computing environment over the past 19 years. Modern anomaly detection systems that are based on Denning's premise tend to function poorly in production environments with many false positives and difficulty in tuning, because the underlying premise is wrong for the environment. There are many instances where normal looking events are in fact malicious and there are numerous cases of anomalous looking events that are not malicious. While this is acknowledged by anomaly detection researchers, it is not given as much importance as we believe it deserves.

In this paper, we are not so much proposing a new paradigm for anomaly detection that will radically alter the field and provide solutions to specific problems, but instead are suggesting that by carefully examining the assumptions common to most anomaly detection research, we are re-examining an old paradigm. The main benefit of this approach will be to clarify and precisely define generally accepted concepts found in anomaly detection studies so that future work will be based on more complete knowledge and thus hopefully produce better anomaly detectors. Another benefit of questioning an accepted paradigm is to begin the process of replacing it if necessary in light of recent developments or else reaffirm that the paradigm is still valid in spite of changes within the field. This last point is the main purpose of our paper.

In Section 2 we define for commonly used terms in anomaly detection research. These terms provide the context for our discussion. Section 3 presents a critical review of the literature with the goal of identifying the assumptions underlying anomaly detection systems with an emphasis on network anomaly detectors. We point out that our review is not meant to be comprehensive with regards to anomaly detection research but rather we include papers whose results are most relevant to the discussion. In Section 4 we question the assumptions that we identified in Section 3, providing material from both within and outside the anomaly detection community to support the idea that these assumptions may be wrong, or should at least be revisited and perhaps refined. We believe that the anomaly detection community needs to recognize the requirements for studying and quantifying what constitutes normal behaviour, malicious behaviour and anomalous behaviour given today's Internet. We provide some concluding remarks and suggestions for future research directions in anomaly detection in Section 5.

2. DEFINITIONS

In order to determine how well anomaly detectors perform, we must first provide a standard set of definitions for each of the key terms. In particular, we define the terms anomalous and malicious.

An anomaly detection system develops a model of normal behavior, and then defines activity which deviates from predictions generated by that model as anomalous. Maxion and Tan [19] define three types of anomalies in their work: *foreign-symbol* anomalies, *foreign n-gram* anomalies and *rare n-gram* anomalies. Foreign symbol anomalies occur the first time a character or item is encountered by the IDS. The appearance of a previously unseen sequence of characters is called a foreign n-gram anomaly. A rare n-gram anomaly occurs when a sequence of characters appears more than once, but below a user specified threshold, such as 5% of the time.

These definitions can be extended to describe network anomaly detectors, however are not sufficient. Many network based anomaly detectors look for changes in the *behaviour* of the network, which is not captured in the definitions by Maxion and Tan. For example, while a particular sequence of n-grams, or packets, might not be rare, a sudden or unexpected increase in the number or rate of occurrences might be considered anomalous. Lakhina *et al.* [13] have identified nine different anomalous behaviours in connection information from network traffic, which they detect using an entropy-based approach. Thus we add to the previous definitions of anomalous: a behavioural change is anomalous if it deviates too much from the baselined activity, where "too much" can be defined by the user (*e.g.*, two standard deviations). Note that we use baseline activity and not normal here. We define a behaviour as some characteristic that can be measured over time.

Unlike anomalous activity, the definition of "malicious" activity is subjective and often site-specific. For example, a government or military site that has a policy against using peer-to-peer services might consider any peer-to-peer activity to be malicious. In contrast, a university or home network might consider such activity to be normal. This interaction with site policy makes testing anomaly detection systems in the context of detecting malicious activity particularly difficult. A system that might work well at one site might generate what another site considers to be a high number of false positives, simply due to differences in policy. Maxion and Tan [19] avoided this issue by focusing specifically on how well detectors performed at detecting anomalies, without regard for what an anomaly might represent. Due to the site-specific nature of malicious behaviour, we do not provide a definition, but note instead that any definition must be developed in the context of a site's security policies.

3. ASSUMPTIONS

Anomaly detection systems begin from the hypothesis that exploitative behavior is quantitatively distinct from normal system behavior. This is the essence of Dorothy Denning's seminal paper on the topic [8], which has since been implemented and expanded upon by numerous researchers over the past 19 years. However, computing and networking environments have evolved considerably since 1987, and the view taken by Denning, whose focus was host-based detection, is perhaps not appropriate to today's networked environments. Regardless, this assumption has formed the basis of most modern anomaly detectors, several of which are discussed further in this section.

While Denning is credited with the initial idea of using anomaly detection to detect exploitive activity, later anomaly detection studies have made several additional assumptions beyond the idea that intrusive activity can be recognized because it is abnormal. We identify nine assumptions, which can be grouped into three broad categories: assumptions about the problem domain, assumptions about the training data and assumptions about the operational usability.

For each assumption we present research that either explicitly or implicitly is based on acceptance of the assumption. We review the results and examine how the particular assumption influenced both detector design and performance.

3.1 Problem Domain

Denning states that her intrusion-detection model is based on the expectation that attacks constitute unusual use of the system and that they are distinguishable from more typical system usage. This hypothesis implies not only that attacks are anomalous ("abnormal"), but also that the anomalous behaviour will be distinguishable from normal behaviour.

The primary assumptions that are made regarding the problem domain are:

- **attacks are anomalous (different from the norm),**

- **attacks are rare, and**
- **anomalous activity is malicious.**

Attacks Are Anomalous
Ertoz *et al.* [10] have developed an unsupervised clustering approach for intrusion detection, called MInDS (Minnesota INtrustion Detection System). Their system analyses connection information derived from flow data, generating clusters of data. An anomaly score is assigned to a connection based on its distance from the cluster and the density of the cluster. They state that "Connections that have high anomaly scores are most likely to be attacks and those with low anomaly scores are most likely to be normal traffic." Thus, like Denning, they assume that attacks are anomalous and that the majority of traffic is benign or legitimate.

Associated with the assumption that attacks are anomalous is the related assumption that attack traffic will be easily distinguishable from normal traffic. Lee *et al.* [14] base their unsupervised learning approaches on an assumption that "attacks are different" from normal data. This implies that attacks can be detected using methods such as outlier analysis, which is one approach used by both Lee *et al.* [14] and Ertoz *et al.* [10].

Attacks Are Rare
Related to the assumption that attacks are anomalous is the assumption that attacks are rare, which is not the same concept. Anomalous implies that an event deviates from the normal or expected behaviour, however it does not necessarily indicate the frequency with which such deviations will occur. In contrast, *rare* indicates that attacks (anomalies) are not common. Lee *et al.* [14] comment that "if the ratio of attacks to normal data is small enough ... the attacks stand out against the background of normal data." (Note that this statement is also related to the previous assumption that attack traffic is distinguishable from normal traffic). The authors use this assumption as the justification for a clustering method, where the smallest clusters are labeled anomalous (and therefore represent intrusive activity). Ertoz *et al.* [10] have also relied on this assumption, stating that "the proportion of network traffic that corresponds to an attack is considerably smaller than the proportion of normal traffic."

Anomalous Activity Is Malicious
In addition to Denning's hypothesis that system exploitation or attacks are anomalous [8], many researchers have also made an assumption of the reverse — that anomalous activity represents attacks or malicious behaviour. In fact, this reverse assumption forms much of the basis of anomaly detection — the assumption that administrators are interested in all anomalous events because it is likely to represent attack activity. While researchers acknowledge that this is not always the case (for example, when discussing their false positive rates), the underlying assumption is still the basis for the approach.

3.2 Training Data

The availability of good data is extremely important to the training of an anomaly detector. If anomaly detectors are trained with data that is not representative of the intended operational environment, the result could be a higher rate of false positives or equally important, a high rate of false negatives.

The common assumptions about the training data include:

- **attack-free data is available,**
- **simulated data is representative, and**
- **network traffic is static.**

Attack-free Data Is Available
Denning [8], when describing her model of intrusion detection, describes the development of usage profiles. She develops statistical models of the observations obtained from audit records, using this as the underlying profiles of expected user and system behaviour. While not explicitly stated, this implies the assumption of attack-free training data. Otherwise, the profiles that were developed would contain attacks as part of the expected user and system behaviour.

Barbará *et al.* [2] describe a testbed that they developed for testing data mining approaches to intrusion detection, called ADAM. This approach requires building a repository of "normal" frequent itemsets, which requires attack-free data.

In later work [3], they describe an approach to detecting attacks based on an unsupervised clustering approach, where they also require a base data set that is composed exclusively of attack-free connections. They generate this set by using only those connections that meet association rules that are found to be common across three different days of data. The implicit assumption behind this approach is that "normal" data will consistently appear across all three days, whereas malicious behaviour (which will be anomalous) will not persist across the three days.

Mahoney and Chan [17] present an approach to detecting rare events in time-series data, called LERAD. Their approach requires two passes of data, the first pass of which requires attack-free network traffic.

The most prevalent approach to dealing with the requirement for attack-free data has been to use a canned dataset for reference, in particular the MIT Lincoln Labs dataset [15]. The Lincoln Labs data set consists of simulated, attack-free or "normal", hosts and network traffic, along with labeled attacks that were generated manually. This data set was developed specifically for testing anomaly detection systems. For example, Mahoney and Chan [17] test their approach, LERAD, in part using the Lincoln Labs data set from 1999. (We note, however, that they also provide results from testing their approach on data gathered from their university network.)

Simulated Data Is Representative
The key assumption made by researchers who used the Lincoln Labs data set for either training or testing anomaly detectors is that this data set is representative of network traffic, and that it is generalizable to other networks.

For example, Ye *et al.* [30] compare the capabilities of several different statistical techniques for detecting intrusions. However, their approach focuses on determining if intrusions can be detected from a single event or if a series of events is required, testing the hypothesis using clean data only from Lincoln Labs and injecting their own attack data.

Sekar *et al.* [26] also used Lincoln Labs data to verify their approach to anomaly detection, which they describe as specification based. In their experiments, they generate state diagrams that represent protocol usage. Network data is used to generate statistical baselines for each state change. Anomalies are detected as statistical outliers of changes in state. While not explicitly identified in the paper, it appears that this approach requires attack-free data for generating the statistics for the state model.

Network Traffic Is Static
An implicit assumption that is made both by those using the Lincoln Labs data set and by those requiring clean data is that the behaviours observed in networks are static. This is an implicit assumption because none of the referenced papers discuss how to perform ongoing training and updating of the anomaly detectors to take into account changes in network traffic composition and concept drift.

3.3 Operational Usability

Denning comments that false alarms "can be controlled by an appropriate choice of statistical model for the activities causing the alarms and by an appropriate choice of profiles." [8] However, this approach is not often used by the anomaly detection community. Rather, the anomaly detection technique tends to be tested and the results presented often with little or no discussion as to how the number of false alarms (false positives) can be reduced. A given false alarm rate may be tolerable in a research setting but completely unacceptable in an actual production environment. Consequently, both the false positive and false negative rates of anomaly detectors should be critically important in determining the effectiveness of an anomaly detector since the operational usability of the detector is affected.

The primary assumptions that are made regarding operational usability are:

- **false alarm rates > 1% are acceptable,**
- **the definition of malicious is universal, and**
- **administrators can interpret anomalies.**

False Alarm Rates > 1% Are Acceptable
One hidden consequence of largely ignoring false positives is an unusually high tolerance for false alarms in the academic literature. Barbará *et al.* [3] tested their approach to creating clusters using the Lincoln Labs data set [15]. Based on one day of data from the training set they found that a threshold of 8.5 resulted in a detection rate of 99% and a false alarm rate of 4%. Using the test data from the 1999 DARPA evaluation, they found that their best threshold was 7.0, which resulted in a 99.9% detection rate and a 14% false alarm rate. This result is somewhat difficult to interpret as the unit of analysis is not known. That is, it is unclear whether the 4% and 14% respectively are based on the percentage of packets that were falsely characterized, or the percentage of sessions, etc.

Mahoney and Chan [17] describe the results for their approach, LERAD, in terms of the detection rate when there are only 10 false alarms per day. In this case, their detection rate varied from 40% using TCP data gathered from their university network (given only a small number of known attacks) to 64% on TCP data from the Lincoln Labs data set.

Similar values are described in the literature for host-based anomaly detectors. For example, Ye *et al.* [30] comment at one point that for their decision tree approach "a hit rate of 88.1% brings up the false alarm rate to *only* 4.6%." [Emphasis ours.]

Definition of Malicious is Universal
Another assumption that is implicit to the previous studies is that the definition of malicious activity is universal. That is, there is no discussion on how the true and false positive rates might interact with differing site policies, but rather the unstated belief that all anomalies are potentially malicious and therefore of interest. For example, Kendall [12] provided a taxonomy of attacks, which was later used for the Lincoln Labs data set [15]. This taxonomy included scans and other probing activities, and represents the set of attacks that is of interest to the U.S. military. This has since been assumed to be of universal interest.

Administrators Can Interpret Anomalies
Xu *et al.* [29] present a clustering approach that uses entropy measures across a four-tuple (srcIP, srcPort, dstIP, dstPort) of flow data collected from backbone routers. They demonstrate that traffic clusters largely into three types of behaviour: servers or services, heavy-hitters (such as web proxies, crawlers and NAT boxes), and scanning activity. They highlight that their approach can identify anomalies that consist of unusual changes to common clusters (for example, sudden scans of unusual ports or unusual profiles being generated for popular services), as well as generating clusters of rare behaviours and recognizing behavioural changes in clusters. Determining if a cluster genuinely contains events of interest is an activity left for the administrator.

Ertoz *et al.* [10], recognizing that administrators may have constraints on their time, try to reduce administrator overhead by grouping anomalies together into sets of network events exhibiting similar characteristics, rather than requiring an administrator to examine each anomaly individually. However, regardless of the attempts to reduce the amount of information to be examined manually, there is still an underlying assumption that an administrator will be able to perform a manual investigation of the clusters of anomalies presented.

4. QUESTIONING THE ASSUMPTIONS

In this section we revisit the assumptions presented in Section 3 and critically examine whether the assumptions hold for the network environment.

4.1 Problem Domain

Attacks Are Anomalous

The assumption that attacks are anomalous, which was originally developed for host based anomaly detection, needs to be verified for network environments. For an attack to be anomalous means it must be distinguishable from normal traffic. However, this is not always possible given the variable nature of the network and the ability of attackers to hide their activities.

Many of the anomaly detection studies that utilize network data don't place enough emphasis on how intruders can hide their attacks. Tan *et al.* [27] investigate how intruders hide their activities within normal data by first identifying a detector's blind spots and then changing the attack to fall within those blind spots. While their work dealt with host-based detectors the authors hypothesized that the same technique would probably work for network based detection.

In another study, Handley *et al.* [21] discuss how intruders could hide their activities from intrusion detectors by taking advantage of traffic ambiguities. The attacker's strategy is to not have their attack appear anomalous but simply hide their actions in the variable nature of the traffic and protocol implementations.

Attacks Are Rare

The assumption that attacks are rare is based on of the original host environment which consisted of normal system users engaged in typical use of the system with only an occasional instance of intruder presence. This assumption is false as applied to today's networked environment where there is a much higher probability of intruder activity. With regards to network traffic, there have been numerous studies recording percentages of legitimate and non-legitimate traffic as the Internet has matured which indicate that attacks are not rare assuming we include scanning in the attack set. Several studies are presented that attempt to quantify the percentage of attack data in network traffic.

Yegneswaran *et al.* [31] studied logs collected from over 1600 globally distributed sites and concluded that scans were rapidly increasing with time and generally occur over a massive scale throughout the Internet. They go on to state that "By projecting intrusion activity as seen in our data sets to the entire Internet we determine that there are typically on the order of 25B intrusion attempts per day and that there is an increasing trend over our measurement period." They also state that "Daily intrusion attempts take place on an massive scale - as many as 3 million scans in our logs on a single day.

Another study by Pang *et al.* [22] used three different network telescopes to characterize Internet "background radiation", where background radiation is traffic sent to unused IP addresses. They comment in their introduction that "The volume of this traffic is not minor. For example, traffic logs from the Lawrence Berkeley National Laboratory (LBL) for an arbitrarily-chosen day show that 138 different remote hosts each scanned 25,000 or more LBL addresses, for a total of about 8 million connection attempts. This is more than double the site's entire quantity of successfully-established incoming connections...."

Confirmation that attacks are not rare was also reported by Xu *et al.* [29], who used entropy measures to develop clusters. The authors comment that "A disproportionately large majority of extracted clusters fall into [the scanning] category, many of which are among the top in terms of flow counts". If scanning is considered to be malicious (which they are in Kendall's attack taxonomy [12], for example), then this indicates that it is not anomalous, but rather quite common!

Anomalous Activity Is Malicious

Assuming all (or most) anomalous activity to be malicious might be true for a single system depending on the OS and applications. However, the Internet has a long history of documented anomalies that are not malicious in nature.

In 1992, Bellovin [5] found that odd packets occurred as a result of router and server problems none of which were malicious. Another study from 1990 documented Ethernet anomalies on a computer science department network [18]. None of the anomalies were due to malicious activity but were the results of a broadcast storm, a "babbling" node, a new network protocol and a graduate student project.

More recently, Mahoney and Chan [17] concluded from their study that, "Many of the anomalies detected by LERAD are not due to hostile code, but rather to legal but unusual protocol implementations."

In another study, Barford and Plonka [4] identify three classes of anomalies: *network-operation*, *flash-crowd* and *network-abuse* anomalies. Network-operation anomalies encompass reconfiguration and transient failures in network architectures, while flash crowds are nonmalicious increases in traffic to target sites. Neither of these two classes of common anomalies represent malicious traffic.

In 2005, Lakhina *et al.* [13] identified eight classes of anomalies that they were able to detect by deploying an entropy-based method against the behaviour of connection information (source IP, source port, destination IP, destination port): alpha flows (very large point-to-point data exchanges), denial-of-service attacks, flash crowds, port scans, network scans, outage events, point-to-multipoint connection (such as content distribution mechanisms), and worms. Of these, only four represent potentially malicious activity (denial-of-service attacks, port scans, network scans, and worms), while the rest represent unusual, but legitimate, events and connection activity.

4.2 Training Data

Attack-free Data Is Available

The assumption that there exists attack-free data for training a detector outside of simulated data is not a realistic assumption. As noted in the previous section (see [22] and [31]), network traffic contains a large number of scans, denial-of-service attacks and backscatter, and worm activity. If not careful, this activity will become part of the normal state for an anomaly detector.

Data from a live network was used to validate NATE, a TCP packet header anomaly detector [28]. Before this data could be used for training, it had to be made attack-free. Taylor and Alves-Foss commented that this required a great deal of traffic screening in order to remove scans and other types of probe activity common to traffic traces from live networks.

Simulated Data Is Representative
Simulated data of any type is suspect unless a convincing argument is made that the simulated data truly represents the actual data being modeled. This is not the case with the Lincoln Labs data, which has documented, known problems, particularly with the network portion of the data set [16, 20]. These problems include: a lack of variability in the traffic types, a lack of non-attack anomalies, and unrepresentative traffic volume for the simulated environment [20]. One other problem that is not generally acknowledged is that this data set is old and not representitive of current network traffic condidtions having been generated in 1998 and 9999. Yet, despite the known problems with this data set, it continues to be used as a basis for training and testing intrusion detection systems.

Using only simulated data to test an anomaly detector runs the very real risk that the anomaly detector will not function as well in a real environment. This was confirmed in a study by Mahoney and Chan [16], who attempted to quantify the difference in network anomaly detector performance between exclusive use of the Lincoln Labs data and mixed Lincoln Labs and real network data. They identified "simulation artifacts" in the Lincoln Labs data, which are those attributes that exhibit high variability in real environments but demonstrated a limited range in the Lincoln Labs data. The artifacts included TCP Time To Live window size, TCP options and the client/source IP address range, among others. After mixing in real data, they observed a much higher attack detection rate for several anomaly detectors when trained with the mixed data.

For many researchers the advantages of using a slightly flawed, but well-known dataset to train an anomaly detector outweigh the disadvantages of using that dataset. Among the reasons stated for using a dataset with flaws include availability of a large set of labeled attacks, ability to compare research based on a standard data set, and the lack of problems with privacy issues since the data is simulated [14, 15, 28]. However, there are more serious consequences when the detector being developed depends on the correctness of the data for its inTernal representation of normal versus anomalous.

Maxion and Tan [19] studied the effect of data regularity on anomaly detectors by generating data sets of different regularities as measured by conditional relative entropy and then seeing how well a detector could detect anomalies embedded in datasets of different regularities. The authors show that data regularity greatly affects detector performance and they cautioned against deploying an anomaly detector into environments where regularity differs significantly from that of the training set. As a result of this study, researchers should be cautious when deploying an anomaly detector trained with the Lincoln Labs dataset in a real network environment because the regularity (or variability) differs between these two environments.

Network Traffic is Static
The assumption that network data is even remotely static over both short and long time frames has been negated by multiple traffic studies. Paxson [23] presented statistics of Internet traffic that confirms the huge variability of traffic across sites, over time and by source. According to Paxson, there is no such thing as a typical site because traffic is in continuous flux. kc claffy from the CAIDA project, whose purpose is to model Internet traffic, is also frequently cited with regards to the difficulty in modeling network traffic since it is so highly variable in nature [6].

The fact that Internet traffic is continuously changing in terms of content, volume and percentage of attacks versus legitimate traffic appears to be acknowledged by most researchers engaged in anomaly detection research, as it affects the detector's ability to distinguish attacks. Thus, most detection algorithms appear to account for the dynamic nature of network traffic for performance reasons. However, researchers typically fail to address re-training or recalibration of detectors which is important for detector usefulness over time especially in environments that are fairly unstable. Consequently, we believe that the assumption of static network traffic is implicit in most network anomaly detection research since the authors typically fail to address retraining or update issues for their detectors. This does not appear to be as much of a problem with signature based intrusion detection systems where there is a more obvious path to updates by the creation of additional signatures for new attacks (*e.g.*, in Snort [25]).

Some researchers acknowledge a strong need for updating an anomaly detector over time. Re-training was stressed by Maxion and Tan [19], who discussed the notion that normal tends to drift over time and any system capable of learning normal must be able to track drift. Dorothy Denning also stressed the idea of normal adjustment with time since she advocated a heavier weighting of more recent behavior in her characterization of normal profiles [8].

4.3 Operational Usability

False Alarms > 1 Percent Are Acceptable
The assumption that relatively high percentages of false alarms are an acceptable price to pay for anomaly detection is partially the result of incomplete problem definition. Most of the anomaly detection research community place a greater emphasis on detection than false positive constraint. While some researchers report tradeoffs in detection accuracy versus false positives when setting detection thresholds, these studies are an exception. We believe a greater emphasis should be placed on constraining the false positives in a network environment in order to produce usable anomaly detectors since even a 1% false positive rate can generate thousands of alerts per day depending on the traffic volume. The issues with a high (or even a modest) false positive rate were highlighted in a paper by Axelsson [1], where he discussed the issue of non-attack traffic being $\gg$ than attack traffic causing an unacceptable *number* of false positives, because a small percentage of a large number is still a large number!

Definition of Malicious is Universal
Returning to our discussion on definitions (see Section 2), the definition for malicious activity can potentially vary be-

tween organizations. This can be for two reasons. First, the definition of malicious for any given organization is related to their site policy. Thus activity that might be considered benign in one network (*e.g.*, peer-to-peer (P2P) traffic) could be considered malicious in another, simply because the site policy states that P2P traffic is not allowed on the network. A second reason that malicious activity will vary between organizations is based on the priority that each organization gives to the activity. For example, one organization might consider scanning activity to be malicious and therefore want to record and/or block that activity. However, a second organization might not be interested in such activity, considering it a nuisance at best. Thus anomaly detection systems need to be designed that can classify the type of anomalous activity they are detecting, allowing the end user to then specify which of these activities are of interest and providing the option to ignore all alerts generated by other activities.

Administrators Can Interpret Anomalies
This assumption is not particularly realistic given the amount of work that typically characterises system administrators who manage networked systems. Administrators must constantly patch applications, update firewall rules and manage users with all their inherent human introduced security problems. The assumption that a system administrator has the ability, time or interest to identify unknown anomalies when there are already abundant known threats is not a practical assumption for maintaining an anomaly detector. The on-line journal of Computer Economics [7] report that, "Most IT departments are overworked and understaffed without adequate time to develop adequate security procedures and processes."

5. CONCLUDING COMMENTS

Denning discusses a number of open issues in the conclusion of her seminal paper on intrusion detection models. In particular, she notes:

- *Soundness of Approach* - Does the approach actually detect intrusions? Is it possible to distinguish anomalies related to intrusions from those related to other factors?
- *Completeness of Approach* - Does the approach detect most, if not all, intrusions, or is a significant proportion of intrusions undetectable by this method?

Even 19 years later, the community has not attempted to address these questions, but rather continues to make the same assumptions, and is additionally making these assumptions about network data rather than host data!

In this paper we call into question the assumptions surrounding anomaly detection, with a focus on network-based anomaly detectors. We provide a critical review of the literature, highlighting the assumptions that underlie the various detectors. We identify nine different assumptions that can be classified in three categories: problem domain (*e.g.*, that anomalous behaviour is malicious), training data and operational usability. We go on to discuss the characteristics of network data in relationship to these assumptions, highlighting those assumptions that need to be reviewed in light of current network traffic patterns. We note that malicious and unwanted traffic has become prevalent enough that we believe that the entire field of anomaly detection as applied to networks needs to be reconsidered. Many of the original assumptions are not valid, or at the very least need to be redefined in the context of today's network characteristics.

At a minimum, we recommend:

1. A consideration for whether anomalies are actually what should be detected. That is, a better approach might be to determine what malicious activities we want to detect, and what characteristics of those activities might appear as anomalous, and then focus on detecting those specific activities. This addresses the blind adherence to the idea that malicious and anomalous are somehow equivalent. It also potentially addresses some of the usability issues surrounding false positive rates.

2. A combination of anomaly detection and classification approaches. Anomaly detection alone places a large burden on the administrator of a network, who must then analyse and manually classify each anomaly. The grouping of anomalies (as performed in MInDS [9] for example) is a good start, however if each anomaly could be further classified as to the type of behaviour or anomaly detected, it would allow an administrator to prioritize. Current approaches of prioritizing by how anomalous an anomaly is does not necessarily capture what an administrator might consider to be the more important security events.

3. Testing testing testing! The Lincoln Labs data set [15] was a good idea, however is no longer an appropriate data set. Real network data is required, however it suffers from the lack of ground truth. An alternative, community-based, testing approach might be to find a good data set (perhaps one available through DHS's Predict project) and then have the community use this data set consistently for testing. Results should be published in a public location, allowing the data set to slowly become labeled. This data set should be updated on a yearly basis to reflect new trends in network data, while still allowing access to older sets whose behaviour may be better understood. While this is not an ideal solution, it is at least better than the ad hoc methods currently employed for testing different network detectors.

4. A re-examination of what defines malicious behaviour of interest. For example, Xu *et al.* [29] found that, of three main clusters of network data, one represented scanning activity. Additionally, background radiation [22] is distressingly common. This indicates that certain malicious behaviour may no longer be anomalous, but actually the norm! Perhaps it is time that we applied anomaly detection to the detection of legitimate traffic, filtering it out and leaving the majority for further analysis!

6. ACKNOWLEDGMENTS

The authors would like to thank Bob Blakley for his excellent note taking during the workshop and the other workshop participants for their comments plus the reviewers who took the time to read the work. We also want to thank Dorothy Denning for her help and review of the paper prior to submission.

7. REFERENCES

[1] S. Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3):186 – 205, 2000.

[2] D. Barbará, J. Couto, S. Jajodia, and N. Wu. ADAM: A testbed for exploring the use of data mining in intrusion detection. *SIGMOD Record*, 30(4):15 – 24, 2001.

[3] D. Barbara, J. Couto, J.-L. Lin, Y. Li, and S. Jajodia. Bootstrapping a data mining intrusion detection system. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, pages 421 – 425, Melbourne, Florida, USA, 2003. March 9-12, 2003.

[4] P. Barford and D. Plonka. Characteristics of network traffic flow anomalies. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*, San Francisco, USA, 2001. Extended abstract.

[5] S. Bellovin. Packets found on an internet. Technical report, AT&T Bell Laboratories, May 1992.

[6] K. Claffy. Internet measurement: Myths about internet data. `http://www.caida.org/publications/presentations/Myths2002/`, 2001. Presentation at the 15th Large Installation Systems Administration Conference (LISA 2001).

[7] R. Collette and M. Gentile. Combating back door vulnerabilities in data center procedures. *Computer Economics*, 2006.

[8] D. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, 13(2):222 – 232, 1987.

[9] L. Ertoz, E. Eilertson, A. Lazarevic, P.-N. Tan, P. Dokas, V. Kumar, and J. Srivastava. Detection of novel network attacks using data mining. In *Proceedings of the 2003 ICDM Workshop on Data Mining for Computer Security*, Melbourne, Florida, USA, November 2003.

[10] L. Ertoz, E. Eilertson, A. Lazarevic, P.-N. Tan, V. Kumar, J. Srivastava, and P. Dokas. *MINDS — Minnesota Intrusion Detection System*, chapter 11, pages 199–218. MIT Press, 2004.

[11] S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. A sense of self for unix processes. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy*, pages 120 – 128, Los Alamitos, CA, 1996. IEEE Computer Society Press.

[12] K. Kendall. A database of computer attacks for the evaluation of intrusion detection systems. Master's thesis, Massachusetts Institute of Technology, 1998.

[13] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *SIGCOMM '05: Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 217–228, New York, NY, USA, 2005. ACM Press.

[14] W. Lee, S. J. Stolfo, P. K. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop, and J. Zhang. Real time data mining-based intrusion detection. In *Proceedings of DISCEX II*, pages 89 – 100, June 2001.

[15] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In *DARPA Information Survivability Conference and Exposition*, volume 2, pages 12 – 26, 2000.

[16] M. V. Mahoney and P. K. Chan. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In *Proceedings of the Sixth International Symposium on Recent Advances in Intrusion Detection*, pages 220 – 237, Pittsburgh, PA, USA, September 2003.

[17] M. V. Mahoney and P. K. Chan. Learning rules for anomaly detection of hostile network traffic. In *Proceedings of Third IEEE International Conference on Data Mining*, pages 601 – 604, 2003.

[18] R. Maxion and F. Feather. A case study of ethernet anomalies in a distributed computing environment. *IEEE Transactions on Reliability*, 39(4):433 – 443, 1990.

[19] R. A. Maxion and K. M. Tan. Benchmarking anomaly-based detection systems. In *Proceedings of 2000 International Conference on Dependable Systems and Networks*, pages 623 – 630, June 2000.

[20] J. McHugh. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security*, 3(4):262 – 294, 2000.

[21] C. M.Handley and V.Paxon. Network intrusion detection: Evasion, traffic normalization, and end-to-end protocol semantics. In *Proceedings of the 10th USENIX Security Symposium*, Washington, DC, August 2001.

[22] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, pages 27 – 40, Taormina, Sicily, Italy, October 2004.

[23] V. Paxson. Why understanding anything about the internet is painfully hard. Technical report, UCB Berkely MIG Seminar, April 1999.

[24] V. Paxson and S. Floyd. Why we don't know how to simulate the internet. In *Proceedings of the 29th conference on Winter simulation*, pages 1037–1044, Wisconsin, 1997. ACM Press.

[25] M. Roesch. Snort — lightweight intrusion detection for networks. In *Proceedings of the 13th Systems Administration Conference*, pages 229 – 238, Seattle, WA, USA, November 1999. Usenix Association.

[26] R. Sekar, A. Gupta, J. Frullo, T. Shanbhag, A. Tiwari, H. Yang, and S. Zhou. Specification-based anomaly detection: A new approach for detecting network intrusions. In *Proceedings of ACM Conference on Computer and Communications Security*, pages 265 – 274, Washington, DC, USA, November 2002.

[27] K. Tan, K. Killourhy, and R. Maxion. Undermining an anomaly-based intrusion detection system using common exploits. In *International Symposium on Recent Advances in Intrusion Detection (RAID) 2002*, pages 54 – 73, Zurich, Switzerland, 2002.

[28] C. Taylor and J. Alves-Foss. An empirical analysis of nate: Network analysis of anomalous traffic events. In *Proceedings of the 2002 Workshop on New Security Paradigms*, pages 18–26, 2002.

[29] K. Xu, Z.-L. Zhang, and S. Bhattacharyya. Profiling internet backbone traffic: Behavior models and applications. In *Proceedings of the 2005 ACM SIGCOMM Conference*, pages 169 – 180, Philadelphia, PA, USA, August 2005.

[30] N. Ye, X. Li, Q. Chen, S. M. Emran, and M. Xu. Probabilistic techniques for intrusion detection based on computer audit data. *IEEE Transactions on Systems, Man, and Cybernetics*, 31(4):266 – 274, 2001.

[31] V. Yegneswaran, P. Barford, and J. Ullrich. Internet intrusions: Global characteristics and prevalence. In *Proceedings of the 2003 ACM Joint International Conference on Measurement and Modeling of Computer Systems*, pages 138 – 147, San Diego, California, USA, June 2003.