Micro-Signatures: The Signatures Hidden in Anomaly Detection Systems

Abstract

The field of intrusion detection is divided into signature detection and anomaly detection. The former involves identifying patterns associated with known attacks and the latter involves attempting to learn a 'normal' pattern of activity and then producing security alerts when behaviors outside of those norms is detected. The n-grams methodology has arguably been the most successful technique for anomaly detection (including for network packet inspection).

In this work, we identify a new type of intrusion detection that neither uses typical signatures nor is anomaly based (though it is closely related to both). We generate n-grams from both malicious content and Snort signatures and use sets of these 'micro-signatures' to identify attacks. This micro-signature capability arises implicitly when the training sets for n-gram anomaly detection systems are scrubbed of malicious content and thus is not new. It was added explicitly by the seminal Anagram network anomaly approach, but was portrayed as a minor enhancement and its effect was not evaluated. In reproducing the Anagram results we find that for our data, the micro-signatures provide the vast majority of the detection capability. What appears on the surface to be an anomaly detection approach achieves most of its effectiveness from a (sometimes merely implicit) signature subsystem. We furthermore find that these micro-signatures enable highly effective standalone detection systems as well as hybrid micro-signature/anomaly systems that generalize to multiple attack classes.

Our results thus shed new light into the functioning of n-gram anomaly detection systems, reveal the need to evaluate the micro-signature contribution within n-gram anomaly research, and open a new avenue of research into how to best use micro-signatures in future detection systems.

1. Introduction

The field of intrusion detection has been an active area of research since at least the late 1980's [1] [2] [3] and is divided into two areas: signature detection and anomaly detection. Signature based intrusion detection systems (IDSs) identify patterns associated with known attacks. Anomaly based IDSs attempt to learn a 'normal' pattern of activity and then produce security alerts when behaviors outside of those norms is detected.

The n-grams methodology has arguably been the most successful technique for anomaly detection. In the late 1990's, the use of n-grams was discovered to be useful for host based anomaly detection [4]. N-grams are simply a collection of arrays of length *n* obtained by applying a sliding window of length *n* to whatever activity is being monitored (e.g., system calls) [5]. N-grams were first applied to analyze network payloads in the PAYL model [6] in 2004 but were limited to 1-grams, as the number of different n-grams that can be acquired can approach a^n where a is the number of characters available (e.g., UTF-8 encoding has 1,114,112 code points [7]). In 2006, the seminal Anagram approach for network packet inspection introduced using an *n* value of greater than 1 by discarding frequency information, accepting a small false positive error, and simply storing the set of acquired n-grams in Bloom filters [8].

In this work, we identify a new type of intrusion detection that uses n-grams but is neither anomaly detection nor does it use typical signatures (although it is closely related to both approaches). We generate n-grams from both malicious content and Snort signatures and use sets of these 'micro-signatures' to identify attacks. The micro-signatures are automatically generated and it is groups of signatures that together detect attacks (as opposed to a single signature mapping to a single attack as in most signature based IDSs). We find that these micro-signatures can be used to create highly effective standalone IDSs or can be coupled with n-gram anomaly detection systems for greater detection scope.

We claim to have 'identified' this approach as opposed to 'invented' because it has always occurred implicitly as a hidden sub-system within n-gram anomaly detection. Whenever one scrubs anomaly detection training data of malicious content, a set of n-grams are removed that then get detected as 'novel' (for those n-grams that don't also occur also in some set of benign training data). The very act of cleaning training data implicitly creates and deploys a set of micro-signatures.

Anagram [8] was the first (and only in our literature search) to include the micro-signatures explicitly as a subcomponent (although the work of [9] includes them in also examining the work of [8]). Doing so enabled them to score the micro-signatures differently than the anomalous n-grams whereas in the implicit approach described above they are scored the same. However, the micro-signature contribution to Anagram was portrayed as a modest enhancement to fine-tune the output of an already high-performing system and its effect was not evaluated.

Here, we reproduce the seminal Anagram results for network anomaly detection (using two Anagram style IDSs and a pure anomaly based IDS) and specifically evaluate the contribution of the micro-signature subcomponent vs. the anomaly detection component. We find that on our data, the micro-signature component performs the vast majority of the detection work with the anomaly component providing a minority input.

The discovery of the effectiveness of the micro-signature component then led us to create standalone micro-signature based IDSs with no anomaly component. To our knowledge, such detection systems have never before been created and tested. We find that this approach has higher overall performance in our experiments compared to the Anagram approaches (albeit by a small margin).

This result does not imply a lack of value to anomaly detection. On the contrary, when the malicious content used to train the micro-signature based component in Anagram did not map well to the set of attacks to be detected in the test set, we see the contribution of the anomaly portion becoming predominant and the micro-signatures playing a supporting role (the converse of what we normally found).

In summary, the primary findings of this paper are the following:

1. N-gram based anomaly detection systems necessarily have two detection components, an anomaly detector and a micro-signature detector.

2. The micro-signature detector is a new type of intrusion detection, mixing anomaly and signature based techniques (n-grams, automatically generated signatures, groups of signatures collectively identifying attacks).
3. The micro-signature component performs the vast majority of the detection work in our reproduction of the seminal Anagram experiments.
4. Micro-signature and n-gram anomaly based system effectively co-exist with each component providing majority input in situations where their relevant strengths apply.
5. Micro-signatures can be used to form highly effective standalone IDSs.

These findings impact the area of intrusion detection in the following ways. First, we have increased the understanding of how n-gram anomaly systems works by identifying the two detection components (anomaly and micro-signature). Second, our results indicate that future n-gram research needs to separately calculate the contribution of the anomaly portion vs. the micro-signature portion to provide accurate measurements of performance (e.g., what appears to be an enhancement to anomaly detection in some research may in reality be due simply to the use of a more exhaustive set of micro-signatures). We question how much of published 'anomaly' detection research really is primarily signature based (we truly don't know). Third, we have opened up a new avenue of research (that is neither anomaly detection nor typical signature detection) in how to best optimize and deploy micro-signatures based IDSs.

The rest of this paper is organized as follows. Section 2 discusses our data sets. Section 3 described how we construct the IDSs. Section 4 provides our results and section 5 is a higher level discussion of these results. Section 6 focuses on the impact of our results to the research community while section 7 provides a list of experiments available for future research. Section 8 discusses related work and section 9 concludes.

2. Data Sets

We used three data sources to create training and test sets:

1. From an operational web server, we collected 106,472,207 port 80 requests over 294 hours.
2. From a combination of scanning, vulnerability assessment, fuzzing, and exploit tools that were targeted at a virtual machine running an identical web stack to the operational web server, we collected 393,814 unique port 80 malicious requests.
3. From a recent set of Snort signatures (version 2962 of the community rules) combined with 301 binary malware samples, we collected 24,883,806 bytes.

From these raw data sets, we generate a gold filter, two bad content filters, five normal filters, a web server test set, and a penetration test set (the use of these are explained in subsequent sections). Note that we constructed these filters and test sets following the same process as described in the Anagram experiments that we reproduced [8]. All port 80 data was pre-processed by stripping off IP and TCP headers.

The gold filter is a set of n-grams stored in a Bloom filter that are to represent non-malicious traffic that has been rigorously checked. To create this filter, we used the first 24 hours of the web server traffic after scrubbing it of malicious packets by using automated tools (including both signature-based and anomaly-based tools) as well as human inspection.

The bad content filters are sets of n-grams that represent malicious activity. One bad filter was created from the first 196,907 requests from the exploit tool dataset. The other bad filter was created from the Snort signatures and malware samples. For the Snort signatures, we used all "content" fields at least *n* bytes in length as well as all fixed terms from the "pcre" fields when those terms were at least *n* characters long. Note that generated n-grams were only added to the bad filters if they did not match any n-grams in the gold filter.

The normal filters are sets of n-grams presumed to be non-malicious (to be much larger than the gold filter since the same rigor of checking is not performed). The create these filters, we used the 198 hours of web server requests that followed the 24 hours used for the gold filter (note that 72 hours of this data remain which we use below). To compare training size effectiveness, we created normal filters using 0, 10, 50, 90, 130, 170, and the full 198 hours of this training date. As in [8], we sanitized this training data as follows: we did not insert any n-grams that matched the bad content filter and we did not insert any n-grams from a packet that had 5% or more of its n-grams match the bad content filter.

The Bloom filters used for the above data sets were constructed using a 2^{24} bit index with 3 hash functions per item and using SHA-1 as the hash function, as in [8]. We used an n-gram size of 5 as [8] cited this as being good for 'general' purpose experiments.

The primary test set consisted of 72 hours of unused web server requests. This data was carefully scrutinized using the same method as with the gold filter to label the requests as malicious and non-malicious. 6271 "malicious" packets were found containing a combination of port scans, web server content enumeration scans, SQL injection attacks, and malformed content which appeared to be designed to evoke error messages for service fingerprinting. We refer to this test set as the 'web server' test set.

An additional test set was created from the remaining 196,907 unused port 80 malicious requests taken from the suite of exploit tools. Since this test set consists entirely of malicious requests, it is not used directly. Instead, it is combined with the web server test set to provide what we refer to as the 'augmented' test set.

3. Intrusion Detection System Construction

We used the gold filter, two bad content filters, and seven normal filters along with four different scoring rules to construct a total of 56 different IDSs (1 gold x 2 bad x 7 normal x 4 scoring rules).

To score a particular request, we matched each n-gram against the various filters to produce an ordered tuple (n_1, n_2, n_3) containing a) n_1 as the number of n-grams that matched the normal or gold content filter; b) n_2 as the number of n-grams that matched the bad content filter (referred to as the "micro-signature filter"); and c) n_3 as the number of n-grams that appeared in neither (for the sake of brevity we refer to this last filter as the "novel content filter," however it is never explicitly constructed). It is clear by construction that these three counts are disjoint, their sum is the number of n-grams in the packet. A score for a given tuple is generated by a normalized inner product:

$$S = \frac{\sum_i n_i w_i}{\sum_i n_i}$$

The selection of w corresponds to a particular scoring rule, of which we consider four:

1. The original Anagram scoring rule: $w_1 = 0, w_2 = 5, w_3 = 1$, which we refer to as "Anagram-(0,5,1)"

2. An unweighted version of the Anagram scoring rule: $w_1 = 0, w_2 = 1, w_3 = 1$, which we refer to as "Anagram-(0,1,1)"
3. A scoring rule which considers only n-grams from the bad content filter: $w_1 = 0, w_2 = 1, w_3 = 0$, which we refer to as "Micro-signature-(0,1,0)"
4. A true anomaly scoring rule which scores on never before seen n-grams: $w_1 = 0, w_2 = 0, w_3 = 1$, which we refer to as "Anomaly-(0,0,1)". Note that to avoid deliberately adding bad n-grams to traffic considered 'normal' under this approach, we used an empty bad content filter. Thus the known malicious traffic was not processed, enabling the related n-grams to be detected as 'novel'. This is a naïve approach as described below and is used to fully explore the set of possible scoring classes.

It is worth noting that – ignoring the magnitude of the weights and restricting them to the same sign – there are 8 possible classes of scoring rules, which can be reduced to 4 by symmetry. For instance, the (1,0,1) rule is simply the complement of the micro-signature scoring rule (0,1,0). The symmetric class pairing are (0,0,0)/(1,1,1), (0,0,1)/(1,1,0), (0,1,0)/(1,0,1), and (1,0,0)/(0,1,1). The class (0,0,0)/(1,1,1) is trivial, always returning a constant value. The class (0,0,1)/(1,1,0) is a true anomaly detector that reveals never before seen n-grams. This is a naïve approach because it ignores the distinction between the gold, normal, and bad content filters and we did not expected it to be useful for intrusion detection (however, see our results). Class (0,1,0)/(1,0,1) represents the micro-signature scoring rule and class (1,0,0)/(0,1,1) represents the Anagram scoring rules. We thus evaluate all available scoring classes.

In our empirical work, each of the 56 IDSs is applied against both the web server test set and the augmented web server test set for a total of 112 experiments.

4. Results

We compare our IDSs using the usual Receiver Operating Curves. In particular, we focus on the area under the curve (AUC) to compare true positive performance across a wide range of false positive rates.

Table 1 provides a high level comparison, showing the mean AUC for each of the four scoring rules with 90 hours of training data for the normal filter using both the web server test set and the augmented test set along with both bad content data sets. A value of 1.0 indicates perfect classification (a true positive rate of 1.0 may be obtained with a false positive rate of 0) while a value of 0.5 is the value that can be obtained by random guessing.

Table 1. Mean AUC Across Both Test Sets

Intrusion Detection System	Mean AUC
Anagram-(0,1,1)	0.93
Microsig-(0,1,0)	0.94
Anagram-(0,5,1)	0.91
Anomaly-(0,0,1)	0.74

Note how the Micro-signature approach performed equivalently to both Anagram approaches. This means that using the n-gram signatures alone produces comparable overall performance to using the n-gram signatures in conjunction with anomaly detection. Note that while the performance of the Micro-signature approach slightly exceeds the Anagram approaches here, the point of this research is not to identify a better IDS, but to show three things: 1) how micro-signatures have been providing the majority portion of the performance of Anagram based IDSs relative to our datasets, 2) the process of filtering out bad content from training traffic implicitly creates these micro-signatures and thus their use is almost unavoidable, 3) micro-signatures can be used to create effective standalone IDSs.

The pure anomaly detection approach performed much worse. However, this was not surprising as it was truly just detecting on never before seen n-grams. We did not filter out bad n-grams embedded in the training set as doing so would have implicitly moved such bad n-grams into the novel content filter (filter n_3 in section 3), thereby converted the pure anomaly detector into having similar results as Anagram-(0,1,1), which we already evaluate. Notice how the very act of trying to filter out bad content from the anomaly detection system's training set implicitly creates a micro-signature detection capability (which we find in this research to be extremely powerful and never before analyzed in the literature).

One might be tempted from these results to discount anomaly detection altogether and simply rely on micro-signatures. However, we find that in circumstances where the micro-signatures do not correlate well to the attacks in the test set, that the anomaly portion of Anagram automatically jumps in an plays a majority role in detection. Overall though, our data indicates that anomaly detection provides a supporting role to micro-signature detection, which does the vast majority of the detection work. The existing literature (see [8]) asserts the opposite without ever explicitly analyzing the contribution of the micro-signatures. Note that we aren't claiming that our results generalize here to other data sets, but our counter examples demonstrate the need for research efforts to document the contribution of both subsystems.

In the next two sub-sections, we evaluate the AUC results for the four scoring methods using differing combinations of test and training sets and differing amounts of training data. After about 90 hours of training data, the AUC values remain high and relatively stable with respect to the amount of training data (broadly consistent with the findings of [8]). However at lower amounts of training data the behavior becomes slightly less consistent. The behavior of the models with no training data whatsoever is surprising but illuminating, as we discuss in more detail below.

4.1 Results for the Web Server Test Set

We now look in more detail at the results for the web server test set. In Figure 1, we evaluate IDS performance with varying sized training sets for the normal content filter and use the web server exploit tool training content for the bad content filter.

The Micro-signature approach provides slightly better performance (AUC of .987) than the Anagram combined micro-signature/anomaly approaches. Note that the micro-signature set was trained against a set of web exploits and so the trained signature set is appropriate for the target set of attacks to be detected (web server attacks). This likely explains the high performance of the methods using micro-signatures (the Micro-signature and Anagram approaches).

The anomaly approach did not do any better than random except with 0 training data for the normal content filter. This data outlier reflects an anomaly system where only the gold filter was used as normal and all else was flagged as novel. The gold filter was carefully created by scrubbing it of all malicious n-grams thereby implicitly adding the malicious n-grams to the novel content filter. In this special case, the Anomaly-(0,0,1) then acts as a micro-signature detector and gains an enormous performance enhancement. The Anagram-(0,1,1) acts identically here (with an AUC of .963) but divides the micro-signatures and novel n-grams between two equally weighted sets. The Anagram-(0,5,1) rule has slightly worse performance (AUC of .958) because of the unequal 5 to 1 weighting of micro-signatures and novel n-grams.

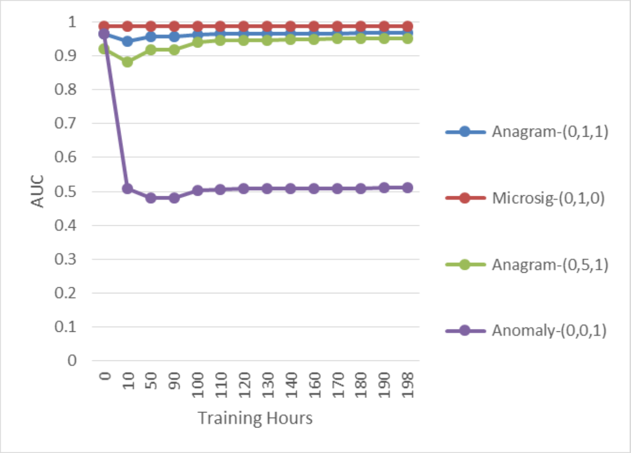

Figure 1. Web Server Test Set with Web Server Exploit Tool Training Content

To look deeper into how micro-signatures contribute to anomaly detection systems in this scenario, we evaluate the relative contribution of novel n-grams vs. micro-signatures within the Anagram-(0,1,1) approach. To do this, we plot in Figure 2 each point correctly classified as 'malicious' by Anagram-(0,1,1) on the x-y plane with the x coordinate being the portion of the score attributable to the micro-signatures and the y coordinate being the portion of the score attributable to the novel content filter. Note that because these two sets are mutually exclusive, all points will lie in the region $\{x > 0; y > 0; x + y < 1\}$. A kernel density estimate to help visualize the distribution of the points is overlaid. The points themselves are plotted with an alpha of 0.05 over the graph; the dashed line indicates equality. In this case (and for all future such plots) we used 130 hours of training data for the normal filter.

A significant number of points in both plots lie along the y=0 line, indicating that none of the n-grams leading to the malicious classification were found in the novel content filter. By contrast, virtually no points lie along the x=0 line. Additionally, the bulk of the density of the distribution lies firmly below the diagonal line, indicating that the majority of the score for the most of the packets was derived from the micro-signature filter, which we thus conclude is doing the 'heavy lifting'.

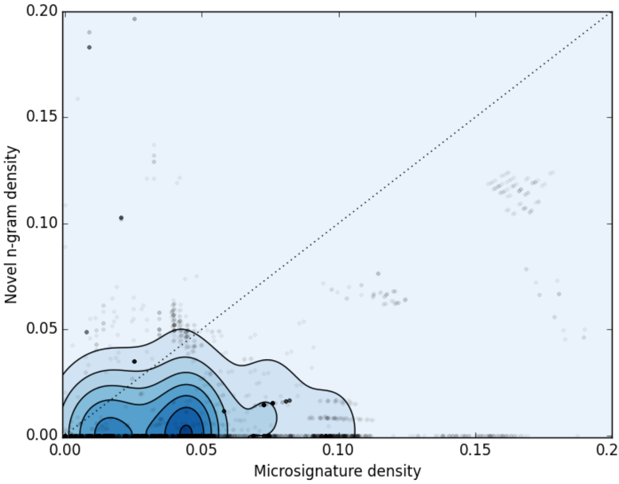

Figure 2. Anagram-(0,1,1) Relative Contribution of Anomaly Detection vs. Micro-Signature Detection for the Web Server Test Set with Web Server Exploit Tool Training Content

We now re-run the same experiment except this time we use the Snort/malware training set (which is not focused on the web traffic being tested) for the micro-signature filter. Figure 3 shows the results. Note the degraded performance of the Micro-signature approach, apparently due to the signature set not aligning as well with the set of attacks to be detected.

However, the Anagram approaches also suffer degraded performance. In part, (shown below) this is because they also rely heavily on the micro-signature filter. But also, consider the consequence of having an ineffective micro-signature filter during training of the anomaly detection capability. If our micro-signatures alone cannot already detect attacks in the training set, then the normal content model that will be constructed will inevitably contain some traffic from malicious packets. This in turn will lead to similar malicious packets being judged to be "more normal" which in turn will lead to a higher rate of false negatives.

Again, the outlier point with 0 training data for the Anagram and Anomaly approaches demonstrates the strength of the micro-signatures. During training, the gold filter was scrubbed of malicious web server traffic n-grams and these micro-signatures were implicitly added to the novel content filter enabling the Anagram and Anomaly approaches to act to a large degree as signature systems.

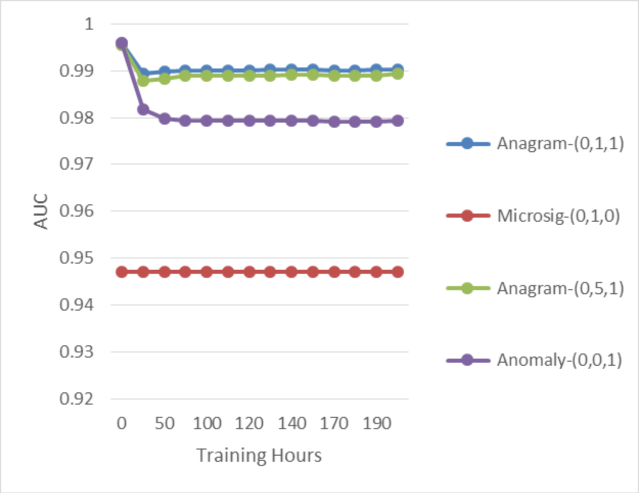

Figure 3. Web Server Test Set with Snort/Malware Training Content

Figure 4 shows the relative contribution of novel n-grams vs. micro-signatures within the Anagram-(0,1,1) approach for this experiment. While the micro-signatures are still doing the majority of the detection work and there are many points on the x-axis (denoting no contribution by the novel component), note that the novel n-grams contribute more than in Figure 2. We attribute this to the anomaly system helping out more when the generated micro-signatures are less appropriate for the attack domain to be detected.

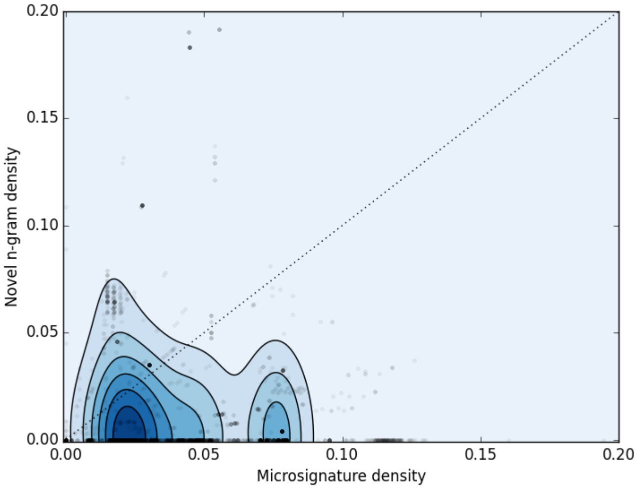

Figure 4. Anagram-(0,1,1) Relative Contribution of Anomaly Detection vs. Micro-Signature Detection for the Web Server Test Set with Snort/Malware Training Content

4.2 Results for the Augmented Test Set

We now look in detail at the results for the augmented test set. Recall that this test set is the web server test set augmented with an additional 196907 malicious requests generated from exploit tools. Given that only 6271 malicious requests were in the web server test set, the overwhelming majority (97 %) of the malicious requests in this augmented test set came from the exploit tools.

In Figure 5 we see that the performance of the Micro-signature and Anagram approaches are similar to that with the web server test set. However, the anomaly approach has improved from performing randomly to obtaining an AUC of almost .98. This increase is explained by noting that the Anomaly approach was only exposed to 6271 malicious requests during training which was not sufficient to prevent it from detecting the 196907 exploit requests as novel (i.e., the micro-signature sets were sufficiently distinct). We posit that had the n-grams generated from the malicious requests in the training data provided more coverage of the malicious requests in the test set, the performance of the anomaly system would have been much worse.

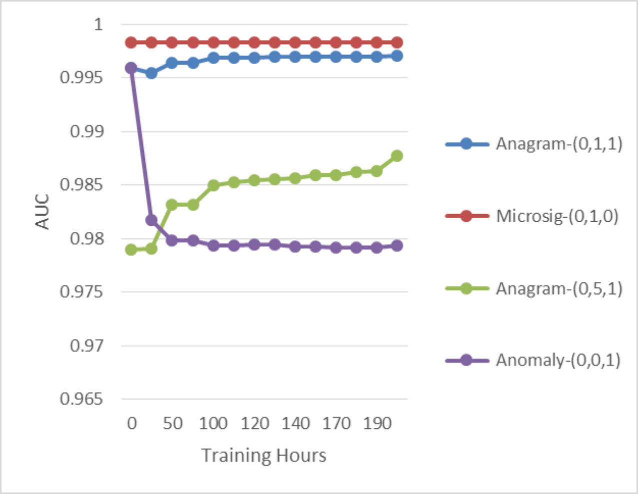

Figure 5. Augmented Test Set with Web Server Exploit Tool Training Content

Figure 6 shows the relative contribution of novel n-grams vs. micro-signatures within the Anagram-(0,1,1) approach for this experiment. Note that to permit a clear visualization of the contours we have omitted the individual points. Here the training set for the micro-signature filter most closely matched the malicious requests in the test set (recall that the web server exploit tool malicious requests were divided equally into a set used for training and a set used for testing).

Note that the novel n-gram density is extremely low compared to the other plots and the micro-signature density is so high that we had to change the x-axis scaling compared to the other graphs just to make the data visible. This can be explained by noting that the micro-signature filter so closely covered the malicious requests in the test set that there were few unmatched malicious n-grams left to label anomalous. We will see the reverse phenomenon happen in the next pair of figures where the micro-signatures do not correspond well to the malicious requests in the test set.

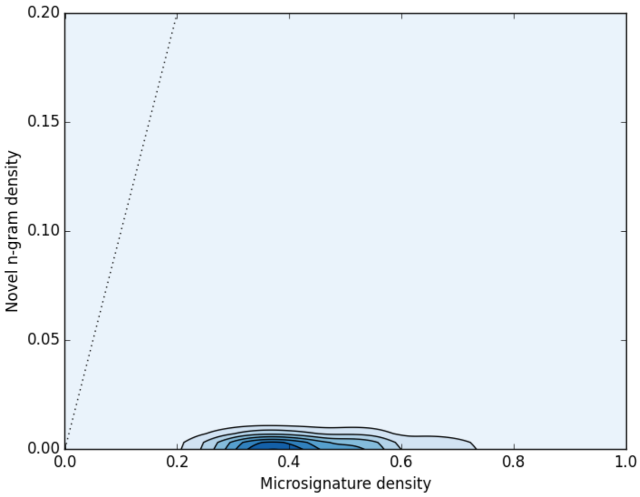

Figure 6. Anagram-(0,1,1) Relative Contribution of Anomaly Detection vs. Micro-Signature Detection for the Augmented Test Set with Web Server Exploit Tool Training Content

We now re-run the same experiment except this time using the Snort/malware training set for the micro-signature filters. Figure 7 shows how for the first time in our experiments, the Micro-signature approach performs worse than the other approaches, albeit by a small margin (note the scaling). The AUC difference between the top performing Anagram-(0,1,1) and the Micro-signatures at 90 hours of training is just .043 and the Micro-signature approach still achieves an AUC of .95. The drop in performance is analogous to the Micro-signature performance drop from Figure 1 to Figure 3. As in this former case, our analysis indicates that the reason is that the Snort/malware training set is less suitable for generating signatures for web server attacks than the web server exploit tool training set.

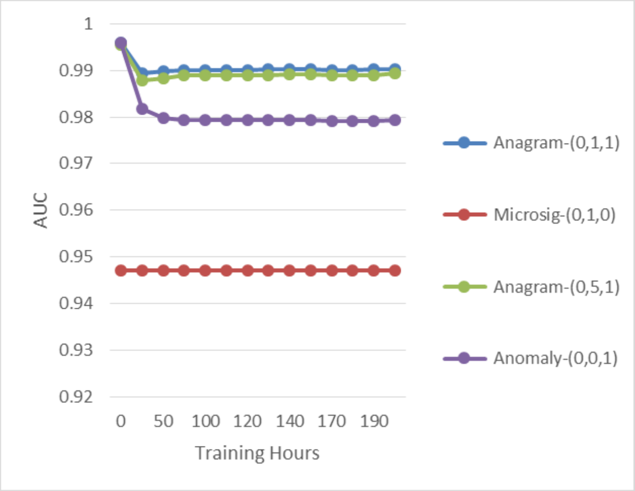

Figure 7. Augmented Test Set with Snort/Malware Training Content

As done previously, we now plot in Figure 8 the relative performance of the novel n-grams vs. the micro-signatures for the Anagram-(0,1,1) approach. We see for the first time the novel n-grams playing a larger role than the micro-signatures. This effect (the converse of that shown in Figure 6) was expected as the micro-signatures generated from the Snort/malware training set poorly matched the test set of malicious web server requests. As a result, the detection burden automatically shifted to the novel n-grams which demonstrates the flexibility of the hybrid micro-signature/anomaly capability within the Anagram approaches. Note, however, that even here the micro-signatures do play a significant role whereas in the converse case of Figure 6, the novel n-grams played almost no role (note the difference in the y-axis scaling of both plots).

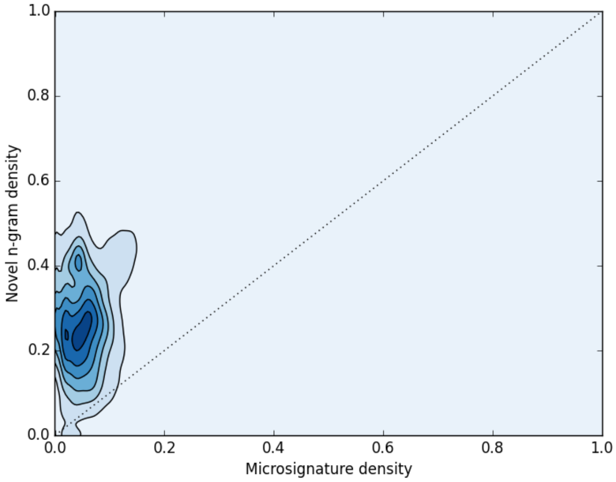

Figure 8. Anagram-(0,1,1) Relative Contribution of Anomaly Detection vs. Micro-Signature Detection for the Augmented Test Set with Snort/Malware Training Content

5. Discussion

Micro-signatures are not a new discovery (having been included within Anagram in 2006), but they were seen as a minor contributor and were not separately evaluated.

Our work reproduces and expands upon the seminal Anagram experiments in [8]. We show that the Anagram approach is clearly effective for HTTP requests. More significantly, we provide the first study that analyzes the contribution to detection made by the subcomponents of Anagram (separating out the anomaly portion from the micro-signature portion). Quite surprisingly, we find that for our data the micro-signatures portion contributed much more to the detection capability than the anomaly portion. This means that, relative to our datasets, the seminal Anagram anomaly detection system that proved the usefulness of n-grams for network packet inspection achieves the majority of its effectiveness from a subsystem that is effectively signature based.

However, this signature based subsystem is very different from typical signature based systems. The signatures are automatically generated from known malicious packets and are very small in size (sets of 5 characters in our experiments). It is the presence of groups of signatures that are indicative of an attack, not just single signatures as is the case with standard signature based IDSs. Because each signature is less focused on a single attack, the signatures appear in our data to generalize to new attacks within the same attack class.

Interestingly, the use of micro-signatures is almost unavoidable for n-gram based anomaly detection systems. Any time training data is filtered of malicious content, the filtered n-grams (unless they also appear elsewhere within unfiltered non-malicious training content) are implicitly forced into the novel content filter. Virtually all n-gram based anomaly systems then benefit from micro-signatures without ever explicitly using them. More thorough and accurate scrubbing of the training data will produce more thorough and accurate micro-signatures. Many papers that have focused on "anomaly detection" using training data that has been scrubbed of malicious content (virtually all of them) have – in effect – been relying heavily on signature-based methods despite being termed "anomaly detection". We make no claim that for other n-gram anomaly systems and datasets we will see the same relative contribution of the components (although we suspect this to roughly true), but one major point of our research is that the relative contribution is important and should be measured.

Another new discovery of this work is that the micro-signatures can be effective on their own, apart from being coupled with an anomaly detection system. Surprisingly, they can function better than the Anagram hybrid micro-signature/anomaly method. That said, we showed how in cases where the attacks in the training set do not correspond well to the attacks in the test set, the anomaly portion of the hybrid approaches kicks in to boost the performance above that of the micro-signatures alone. This leads to the observation that hybrid systems using both micro-signatures and anomaly approaches provide a broader scope in detecting varying classes of attacks.

Another unexpected result was that at 0 hours of training data for the normal filter the anomaly systems performed extremely well. In fact, they often performed the best with 0 hours of training data for the normal filter. We explain this by noting that with 0 hours of training data for the normal filter, the anomaly algorithms are only using the n-grams from the gold filter that were taken from 24 hours of highly scrubbed web server requests. This scrubbing implicitly created micro-signatures that enabled the high detection rate. Our conclusion here is that a smaller amount of carefully scrubbed training data can create a more effective hybrid micro-signature/anomaly detection system than one with a larger amount of less carefully scrubbed data.

As a final issue, we consider the resilience of micro-signatures to evasion attacks. In particular, the normalization to packet length in our Micro-signature approach could lead to an evasion attack where a malicious packet is stuffed with a lot of normal data; this "content mimicry" attack is considered within the original Anagram paper, where it is addressed via subsampling of the packet payload [8]. While the mimicry resistant approach suggested in the original Anagram paper will likely not be as effective for micro-signatures, another potential avenue for handling content mimicry might be through not normalizing the micro-signature counts to packet length. Not shown in this paper are results which find that this idea is effective, but has worse performance than normalized micro-signatures.

6. Impact of Results

How do our results impact the field of intrusion detection? This is an especially valid question since micro-signatures are already being used, albeit unknowingly in virtually all n-gram based anomaly detection. In addition, they were even used explicitly in Anagram, although not evaluated for effectiveness and assumed to be a minor contributor.

One answer is that our results provide us a new understanding of how n-gram anomaly detection functions. We now understand that n-gram anomaly detection systems almost unavoidably contain a signature component (whether realized implicitly or explicitly). When cleaning training traffic of malicious content, if the related n-grams are stored in a bad content filter then they can be used explicitly for micro-signature detection. If they are not stored then the related micro-signatures become implicitly added to the novel content filter. Note that this novel content filter is not usually explicitly created but is the set of n-grams not present in the 'normal' traffic filters. If an n-gram anomaly system attempts to avoid using micro-signatures by not cleaning the training traffic, the performance declines drastically as was seen for our Anomaly-(0,0,1) approach. One could argue that instead of scrubbing malicious data from the training set, that a natively clean training set could be provided to avoid creating micro-signatures. A problem with that approach is that such natively clean data sets are usually created in a laboratory setting and often don't represent the variety seen 'in the wild' (thus much real normal traffic will be considered novel). Finally, we see no compelling reason for n-gram anomaly detection systems to attempt to avoid micro-signature use given there benefit. In our data they provided the majority of the detection capability to the hybrid micro-signature/anomaly detection approaches.

The realization that n-gram anomaly detection invariably contains two components impacts how future n-gram approaches should be measured; the contribution of the micro-signature and anomaly components should be explicitly measured. By doing this, researchers will be able to discover whether or not their new technique is an advance in anomaly detection or that it simply uses a better or more focused set of micro-signatures. Measuring the relative contribution of the two sets is not hard, but it does require the researchers to keep track of the filtered n-grams during the process of cleaning the training sets of malicious data (only including those not found normal in other benign training data). Another reason to measure this in future research is to determine the overall contribution of the two components over more varied sets of data. Our experiments showed the micro-signature contribution dominating for our web traffic dataset. It is not yet known whether or not this result generalized to other data sets. Our current hypothesis based on the results of this paper, completely counter to that of the current understanding, is that n-gram anomaly detection is primarily a signature based approach that is only augmented by anomaly detection. We could be completely wrong, however, finding the answer is important regardless of the discovered result. Only a collection of future studies in a variety of domains will determine this overall trend.

Lastly, our results impact the field of intrusion detection by opening a new avenue of research into a new type of intrusion detection (albeit one closely related to both anomaly and traditional signature detection). This includes exploring using micro-signature IDSs as standalone systems as well as in hybrid systems that combine micro-signature and anomaly detection. While micro-signatures have been implicitly used since n-gram anomaly detection was developed, they have never been carefully studied. By deliberately focusing on their development, we will see how far the approach can be optimized and hope that it will lead to deployable systems. N-gram anomaly detection, despite its success in research circles, has not (to our knowledge) been widely deployed commercially due to unacceptable false positive rates. Perhaps the micro-signature system, being closer to traditional signature based approaches but using n-grams like anomaly systems, will have greater operational applicability and be a stepping stone towards enabling the enterprise deployment of anomaly based approaches.

7. Future Work

This section contains a variety of ideas for additional research in this area. The authors do not have the resources to explore all of these and encourage the community to help fully develop this new research area.

Future work should explore how to most effectively use micro-signatures and how to obtain the best accuracy. The various parameters that can be set for the micro-signatures, including the length of the n-gram used, the parameterization of the Bloom filter (or other data structure), and methods for selecting the threshold parameter in the absence of extensive validation data, all require further study. Future work should explore how well micro-signatures generalize to different types of attacks and never before seen attacks (both within a specific attack class and between different attack classes). The possibility of combining micro-signatures covering different protocols within a single, larger Bloom filter should be explored. Another possible experiment is to create micro-signatures from standard IDS signatures and compare their performance (we expect the micro-signature variant to generalize while the standard signatures will not). A further study can evaluate the extent to which a group of micro-signatures can hinder an attacker from creating variations of attacks that evade current signature sets.

Future work should also be conducted in how to best leverage micro-signatures within n-gram anomaly detection systems. Previously, the micro-signatures were present and used in such systems. Now that we know of their presence, we can research how to best optimize their use in conjunction with the anomaly detection component. One area is to determine the optimal scoring weights for micro-signature and anomaly n-grams for different data sets, or examine alternate ways of deriving more expressive features from them. We saw in our work that the Anagram weighting of 5 for micro-signatures and 1 for novel n-grams performed worse than an equal weighting but no further work was done in this area. Another important step in researching hybrid micro-signature/anomaly systems is in confirming or refuting our conjecture that n-gram anomaly detection in general is primarily a signature based approach. This will likely need to be done by many researchers in different areas testing their unique datasets. While we would have liked to do that within this research, such a breadth of test data is not available to us or (to our knowledge) any single group of researchers. The work of [9] has examined several features of the distribution of n-grams for "normal" packet content that give an indication of how effective n-gram methods are likely to be on such content; it seems worthwhile to explore methods to adapt their measures to malicious content.

Lastly, network forensic related studies should be conducted on how to link micro-signature detected attacks with the relevant source material. This would include identifying the relevant portions of the flagged packets and/or a set of simple signatures (on which the matched micro-signatures were based).

8. Related Work

Given that our work is the first study on micro-signatures, for this related work section we focus on references to n-gram anomaly detection and more general challenges to anomaly detection in the field of machine learning.

The difficulty of applying machine learning in general to intrusion detection is discussed by Sommer and Paxson [12]. They point out several features of intrusion detection problems that make it difficult to successfully apply machine learning. In particular, the rareness of attacks, the high cost of diagnosing detected attacks (particularly when there is a mismatch between the information that a machine learning system provides the analyst and the way in which an analyst diagnoses an event), and the complexity of the input data all mean that machine learning IDS solutions must achieve extremely low error rates on extremely complex problems to be operationally effective. A more probabilistic argument is made in [10] in terms of the base rate fallacy. Nevertheless, multiple examples of anomaly-based and unsupervised network intrusion detection methods can be found in the literature.

One of the earliest n-gram approaches is that of the PAY-L system [6], which clusters network traffic based on the distribution of 1-grams. The Anagram system [8], which forms the basis of our analysis, extends the length of the n-grams to between 5 and 9, while also addressing the issue of "content mimicry". In perhaps the most general case, the issue of anomaly detection via n-grams in non-textual, binary protocols is considered by Hadžiosmanović et al. [9], building on the work of [6] and [8]. This work examines classifiers that make no use of any protocol-specific domain knowledge, and concludes that n-gram based methods generally perform poorly for binary protocols, with an unavoidable tradeoff between high detection rate and low false positive rate. This poor performance relates directly to the 'variability' of the normal traffic (more precisely, the degree to which the n-grams appear to be sampled approximately uniformly from the space of all possible n-grams in normal traffic). While they do not specifically address compressed or encrypted protocols, it seems clear that these will have similar issues. More recently, the work of [9] explores various statistical measures relating to the distribution of n-grams, and relates these measures to the performance of n-gram based supervised and unsupervised classifiers; their work emphasizes machine learning aspects more heavily than our basic analysis, and uses richer feature vectors and more sophisticated classifiers. Similarly to the clustering described in [6], the work of [13] examines the use of a self-organizing map for on-line clustering of packets.

Domain-specific knowledge, in the form of partial parses of protocols, can be used to extract more specific sets of features that help in the identification of anomalous content. In Robertson et al. [14], for instance, web requests are processed by specializing to particular web components, and then extracting key-value pairs from the URIs specific to those components. They learn specialized models (such as simple regular expressions, often simply representing the allowed characters) conditional on each field and component – in effect learning a mixture of site-specific 'sub-protocols' within HTTP. Guangmin [15] performs similar tokenization for use in an artificial immune system model. Ingham et al. [16] attempts to learn deterministic finite automata (DFAs) for normal HTTP traffic while detecting, parsing, and transforming known features (such as email addresses) in order to control complexity. The high degree of structure in the underlying grammar (HTTP) combined with the generally limited character set all contribute to the ability of such systems to be effective. However, these systems are also highly specialized to their particular domain of application and so cannot extend to more general intrusion detection scenarios.

Finally, as machine learning techniques have developed, anomaly-based IDS work has kept pace. More advanced approaches to the problem include that of Gornitz et al. [17]. Here, active learning is used to request that specific packets be labeled by an outside mechanism (e.g. a human analyst) thus maximizing the discriminative power of the learning algorithm within a limited budget of time and effort. While such systems do require more resources to train initially, they typically result in significantly improved performance over purely unsupervised systems. The use of the bad content filter in the Anagram system [8] may be viewed as a non-active, simplified version of this semi-supervised approach.

9. Conclusion

In reproducing the seminal Anagram research for network anomaly detection, we have identified the important role of micro-signature based intrusion detection. We explored how micro-signature detectors are a new type of intrusion detection, mixing anomaly and signature based techniques (n-grams, automatically generated signatures, and groups of signatures collectively identifying attacks). We furthermore discovered that n-gram based anomaly detection systems necessarily have two detection components, an anomaly detector and a micro-signature detector. We found that for our data the micro-signature component performs the vast majority of the detection work but that the anomaly detection component is still important and a significant contributor. On that point, we find that micro-signature and n-gram anomaly based systems effectively co-exist with each component providing majority input in situations where their relevant strengths apply. Finally, we discover that micro-signatures can be used independently to form highly effective standalone IDSs.

Our discoveries are important in several areas. From a foundational point of view, they provides us a new understanding of how n-gram anomaly detection functions. From an anomaly detection research point of view, they leads us to recommend that all future n-gram anomaly detection research calculate the relative contribution of novel n-grams vs. micro-signatures in order to accurately measure the effectiveness of the anomaly detection. From an operational point of view, they lead us to investigate how to best deploy micro-signatures to augment existing intrusion detection systems. Overall, our results provide the initial discovery of a new area of intrusion detection that is neither standard signature detection nor anomaly detection, opening up a new avenue for IDS research.

10. Works Cited

[1] S. E. Smaha, "Haystack: An intrusion detection system," in *Aerospace Computer Security Applications Conference*, 1988.

[2] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. 2, pp. 222-232, 1987.

[3] H. S. Vaccaro and G. E. Liepins, "Detection of anomalous computer session activity," in *IEEE Symposium on Security and Privacy*, 1989.

[4] S. Forrest, S. Hofmeyr and A. Somayaji, "Computer immunology," *Communications of the ACM*, vol. 40, no. 10, pp. 88-96, 1997.

[5] D. Damashek, "Gauging similarity with n-grams: language independent categorization of text," *Science*, vol. 267, no. 5199, pp. 843-848, 1995.

[6] K. Wang and S. J. Stolfo, "Anomalous payload-based network intrusion detection," in *Recent Advances in Intrusion Detection*, Heidelberg, 2004.

[7] "The Unicode Standard Version 6.0- Core Specification," February 2011. [Online]. Available: http://www.unicode.org/versions/Unicode6.0.0/ch01.pdf.

[8] K. Wang, J. J. Parekh and S. J. Stolfo, "Anagram: A content anomaly detector resistant to mimicry attack," in *Recent Advances in Intrusion Detection*, Heidelberg, 2006.

[9] D. Hadžiosmanović, L. Simionato, D. Bolzoni, E. Zambon and S. Etalle., "N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols," in *Research in Attacks, Intrusions, and Defenses*, 2012.

[10] S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection," *ACM Transactions on Information and System Security* , vol. 3, no. 3, pp. 186-205, 2000.

[11] R. Chang, R. E. Harang and G. S. Payer, "Extremely Lightweight Intrusion Detection (ELIDe)," Army Research Laboratory, 2013.

[12] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Security and Privacy* , 2010.

[13] D. Bolzoni, E. Zambon, S. Etalle and P. Hartel, "Poseidon: A 2-tier anomaly-based intrusion detection system," arXiv preprint cs/0511043 , 2005.

[14] W. Robertson, G. Vigna, C. Kruegel and R. A. Kemmerer, "Using generalization and characterization techniques in the anomaly-based detection of web attacks," in *NDSS*, 2006.

[15] L. Guangmin, "Modeling Unknown Web Attacks in Network Anomaly Detection," in *Third International Conference on Convergence and Hybrid Information Technology*, 2008.

[16] K. L. Ingham, A. Somayaji, J. Burge and S. Forrest, "Learning DFA representations of HTTP for protecting web applications," *Computer Networks* , vol. 51, no. 5, pp. 1239-1255, 2007.

[17] N. Görnitz, M. Kloft, K. Rieck and U. Brefeld, "Active learning for network intrusion detection," in *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*, 2009.

[18] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," 2000.

[19] V. Paxson, "Bro: a system for detecting network intruders in real-time," *Computer networks* , pp. 2435-2463, 1999.

[20] M. Roesch, "Snort: Lightweight Intrusion Detection for Networks," in *LISA*, 1999.

[21] K. Rieck and P. Laskov, "Detecting unknown network attacks using language models," in *Detection of Intrusions and Malware & Vulnerability Assessment*, 2006.

[22] K. Rieck, P. Laskov and K.-R. Müller, "Efficient algorithms for similarity measures over sequential data: A look beyond kernels," *Pattern Recognition*, pp. 374-383, 2006.

[23] C.-C. G. F., A. Stavrou, M. E. Locasto and S. J. Stolfo, "Adaptive anomaly detection via self-calibration and dynamic updating," *Recent Advances in Intrusion Detection*, pp. 41-60, 2009.

[24] R. Perdisci, D. Ariu, P. Fogla, G. Giacinto and W. Lee, "McPAD: A multiple classifier system for accurate payload-based anomaly detection," *Computer Networks* , vol. 53, no. 6, pp. 864-881, 2009.