ALDR: A New Metric for Measuring Effective Layering of Defenses [*]

Nathaniel Boggs
Department of Computer Science
Columbia University
boggs@cs.columbia.edu

Salvatore J. Stolfo
Department of Computer Science
Columbia University
sal@cs.columbia.edu

Abstract

Attackers continually innovate and craft attacks that penetrate existing defenses. New security product purchasing decisions are key in order to keep organizations as secure as possible. Current information available to inform these decisions is often limited to individual security product detection/blocking rates for some test set of attacks. Actual security performance, however, depends on how a security product performs in the context of an organization's existing security products. Even a security product that tests well on its own may be completely redundant when deployed into an existing environment. We propose a new metric that measures the total security granted by a combination of security products. Also, this metric makes the computation of the added benefit of an additional security product easy. We take the results of each individual security product parsing a certain data set and then, take the union of the results of all security products deployed at that organization. Our metric is the attacks in this union divided by the total attacks in the data set or, in other words, the total detection rate achieved by the whole system. This metric can be computed using existing evaluation techniques and provides a more accurate overall picture of the security posture of an organization as well as a way to measure the real contribution of a specific security product in the context of other security layers.

1 Introduction

Purchasing decisions for selecting new security solutions currently rely on vendor reputation and support, price, and a comparison of security product effectiveness. To gauge security product effectiveness users have to wade through unstandardized vendor claims, vendor sponsored evaluations, and other marketing hype. At best, new security products will be fairly evaluated by a third party with a particular data set designed to be representative or by the purchasing organization with in house data, but without a good view to establish ground truth. The metrics generated even by the best current tests fail to measure the gain of attack detection using the new security product with regard to the existing layers of security. This leaves those making these key purchasing decisions woefully uninformed as to how each potential security solution will actually affect their organization's overall security.

We propose a new metric that takes all of an organization's layers of security into account and that is computable with little modification to existing testing techniques. Our idea is to take a set of attacks and use the different aspects of those attacks to test appropriate layers and track which attacks are detected by each security product. An attack can be made up of many components. For example a single attack could use a malicious domain, web exploit code, and a host side trojan. Since different layers of security often detect an attack based on a single one of these components, in order to see the true overlap and total detection of a system we need to identify which of these components are part of the same attack. By linking the attacks across layers, we can measure how a set of security products detect attacks as a whole rather than just measure one layer in a

[*] This material is based on research partially sponsored by the National Science Foundation (NSF) under CCF grant 0950373. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF

void. For example, if an attack has a drive-by download and a malware Windows Portable Executable file component we can test URL reputation systems with the drive-by download site data and antivirus products with the Windows PE file. Additional layers such as Intrusion Detection Systems or Network Based Anomaly Detection Systems would be given other appropriate data from the attack such as network packets for instance. The actual tests are the same as existing measurements except that the data sets used are linked instead of selecting a different set of attacks for each layer such as AV versus IDS tests. Once the security products are evaluated at their layer, we can measure the combination of security products in layers by performing a union of the sets of attacks detected by each security product since the same attacks are present in each test. The percentage of attacks in this union out of the total number of attacks is our detection rate metric. Using our metric allows combinations of security products/layers to be more accurately compared, and we can evaluate a new security product based on how much it increases this metric for a particular site with specific layers already in place.

In this paper, we describe our new metric, comparing it to existing measurements, and expanding upon the benefits that using our metric provides the community. Our metric more accurately describes typical modern security infrastructure with its emphasis on layers by evaluating attacks as a whole rather than just individual components of them in isolation. Also, our metric can be expanded to describe false positive rates as well as security system redundancy. Existing measurements that focus on testing security products of the same type against each other only tell us which security product is the best when only one security product is used as an organization's entire defense, which is not the typical practice. Our novel metric, if computed responsibly using good data sets, can provide the security community with an accurate description of an entire defense in depth security setup and the real contribution that each product adds to overall security.

2 Related Work

Previous work in the area of security metrics has often focused on the need for metrics and describing the current best metrics widely used. Andrew Jaquith outlines a number of useful metrics as well as discusses the need for and how to determine utility of security metrics in his book [1]. The Center for Internet Security produces a synthesis of standard metrics and definitions in their CIS Security Metrics report [6]. Most of their metrics deal with measuring internal security over time. They do not address how to measure security products' effectiveness in relation to other security products or with a data set with known ground truth. Our metric begins to fill this void.

Other companies perform security product evaluations by comparing security products to other security products in the same category, such as different antivirus programs, directly using carefully crafted data sets. NSS Labs [3] and similar organizations perform third party security product comparisons by calculating detection and false positive rates with attack data collected from real networks and honeypots. While this is the state of the art in empirical security product evaluations, it lacks contextual information on how security products perform in conjunction with already deployed security products. These security product evaluations also do not measure the total security provided by an organization's many layers of defenses.

Research has also been done to find ways to measure total security for organizations. In [4], the authors describe a model for evaluating network defenses using attack graphs to find the weakest attacker that could likely defeat the defenses. Such models are abstracted to such a point that while useful for evaluating proper placement of layers of security they lack the granularity to suggest empirical measures of the security of a network. On the other extreme, the authors of [5] suggest that models and lab tests of security products fail to accurately represent the real security provided and that clinical trials of security products are needed. These trials would measure the differences in security products deployed to production workstations and networks across large populations and network settings at great cost. Our metric represents a good middle ground between these approaches. Our metric provides an empirical granular measure of individual security products in context and of total performance of security layers with reproducible results at a reasonable cost.

3 Experiment Architecture

We now walk through the steps to compute our proposed metric and more carefully define it. We start with a data set of attacks made up of the data from each attack that each layer of security products could use to detect the attack. For example, some common representations to be archived include source information such as URL and IP, network packet captures, shell code, and Windows PE files. Let A be this set of attacks. Now test security products on the representation of each attack suitable for each security product. Let S_i be the set of attacks that each security product i detects. For example, a host based antivirus product tests each Windows PE file and any attacks it detects are added to its set of detected attacks. A group of security products then have a total number of attacks detected. We take this number divided by the total number of attacks to compute our metric. This represents the effectiveness of a combination of security products in detecting attacks, allowing us to measure the total security of an organization rather than just individual security products.

A is the set of attacks

N is the set of security products tested

S_i is the set of attacks $\{x|x \in A \wedge x$ detected by security product $i \in N\}$

$L \subset N$

$T = \bigcup\limits_{j \in L} S_j$

So T is the set of attacks detected by security products in L

All Layer Detection Rate (ALDR) for security products in L= $\frac{\text{number of elements in } T}{\text{number of elements in } A}$

The framework of our metric is designed to work with all traditional measures of a single security product's effectiveness not just detection rate. Instead of detection rate we can also measure false positive rate, block rate, logging rate, etc just by changing the data set to legitimate traffic in the case of false positive rate or by measuring blocking or logging instead of detection for each security product. This gives us a broader picture of the effectiveness of a group of security products.

An important use of our metric is to compare security products based on the increased protection they afford in the context of existing security products. Answering this useful question now becomes a simple set difference and a bit of arithmetic. Simply compute ALDR for the existing security products, and then compare the ALDR computed for the existing security products with the new potential security product included. Do the same for false positive rates, blocking rate, and any other useful measurements to get a full picture of the tradeoffs for adding the security product. Match these with cost of ownership data to optimize improvement per cost.

Additionally, our metric is easily modified to measure the redundancy of layers. Intuitively redundancy to some degree should be valuable. For instance, two layers could detect the same attack in different ways such that an attacker would have to modify an attack twice to bypass both instead of once to just bypass a single layer. Measuring redundancy is as simple as tracking the number of times each attack is detected instead of just taking the union of the sets. Numbers like how many attacks are detected at least twice could be useful. To be most useful, however, such metrics would have to be calculated only for detection methods that are unlikely to be bypassed by similar attack modification which is beyond the scope of our metric to determine.

4 Benefit

Currently there is no good way to compare the overall security posture of two separate organizations' layers of security. Our metric is the first step towards being able to quantify an organization's security. This would allow compliance with certain standards to be measured in terms of total security rather than the deployment of certain classes of security products. A potentially enormous gap exists between the detection capabilities of security products even in the same category. For example, a leading antivirus vendor with a powerful reputation based system and dynamic analysis can significantly outperform competitors but current compliance rules make little distinction. These rules could be changed to require a

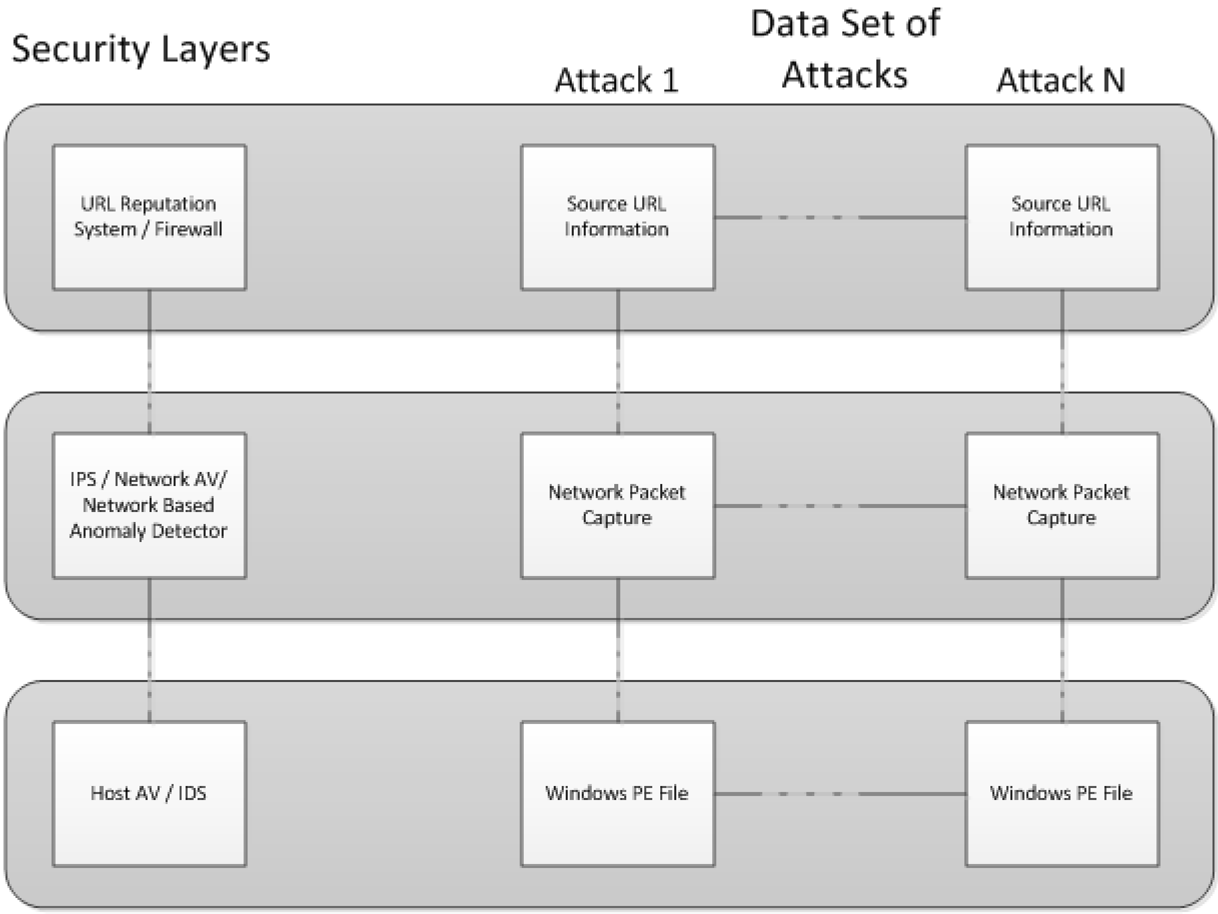

Figure 1: Each layer of security may look at different aspects of each attack, but since the attacks are the same we can see how each attack might be blocked at different layers.

certain detection rate on a representative data set as a compliance standard instead. Using a more holistic metric like ours with a good representative data set as a compliance standard would change compliance to be a step closer to the real goal of making an organization more secure. Such a compliance change could raise the security posture of whole industries making the attackers' job significantly more difficult.

Using our metric we can measure the contribution of each security product to the overall detection rate and false positive rate of an organization's security layers. This will show us which security products are the most vital, but just as importantly, this shows which are the least useful. Some security products could turn out to be completely redundant in that they detect no attacks that the rest of the security layers do not. In fact, a security product could even be detrimental in that it could not detect additional attacks and it could add new false positives. Additionally, this metric could be used to find and remove the security product producing the most unique false positives.

A major benefit to the industry as a whole is that with better metrics security product differentiation is easier and more beneficial. Actual security features will be required to improve this metric rather than just marketing hype so companies are incentivized to improve their security products. Organizations can tell when a security product is redundant so adding real specialized security benefit would be in security vendors' best interest. This could result in new security products that add new layers to current defense in depth strategies or security products good at detecting evasive or zero day attacks with which current security products have trouble.

An organization utilizing our metric would reap significant benefits by being better able to allocate money towards security products. Having actual numbers to measure the added benefit a security product brings would allow an organization to make the most efficient purchase available to them. Without such a metric, an organization is left to depend upon industry best practices or protect against the latest scare. Using our metric organizations would consistently raise the bar for attackers with every new purchase and spur the demand for innovative security products.

5 Data Sets

Metrics relying on real security product performance tests require good representative data sets in order for their application to be useful. Measuring the detection rate and false positive rate of security products for attacks on clean data that are rarely if ever seen will lead to skewed results that fail to represent reality. Proposing new methods of collection or best practices for data set use is beyond the scope of this paper, but we want to touch on certain specific issues that deal with our proposed metric. The data sets being collected for existing tests of security products' detection rates and false positive rates are also suitable for our metric with only a slight increase in the data collected. Linking the representation of attacks across layers will require that data set collection additionally classifies with which overall attack to associate each piece of data with. No additional honeypots or real network data would be required only some further classification so our data set requirements should cost only a small amount more to collect. Obtaining representative data sets remains an open problem, but our metric can be applied as well as existing tests with only a slight increase of difficulty due to the cost of adding some additional labeling to current data sets.

While harder to collect, data sets that include attacks relying on the human element in security could allow our metric to give an even more comprehensive picture of an organization's security. For example, an email phishing data set with information on how often users open attachments or go to websites linked in the emails associated with phishing attacks could be useful in designing and measuring new layers of security. With our metric run on this type of data set the effectiveness of user education with regards to phishing attacks could be compared to an email filtering security product. With some creativity a data set could be collected to allow us to measure social engineering attacks and compare different controls or education. The framework for our metric can easily be expanded to work with all types of attacks provided a proper data set can be collected.

6 Methodology

In creating this new metric for looking at layers of security as a whole, we use the general methodology proposed by Jaquith in [1]. Jaquith proposes five attributes of a good metric: "consistently measured," "cheap to gather," "expressed as a cardinal number or percentage," "expressed using at least one unit of measure," and "contextually specific." Our metric passes all five of these tests for a good metric. With a good testing infrastructure to reliably test new data sets, an organization can "consistently measure" our metric for each security product. Our metric can be tracked over time against new data sets, and we can measure which security products hold up well against newer attacks. Existing testing organizations already collect similar data sets and to measure using our metric simply have to link the different aspects of attacks together across layers such as associating the attack's network packets with its Windows PE file or original phishing email. Since all this data is already being captured, only some linking and reorganization is required so we also pass the "cheap to gather" test. We certainly meet the third and fourth requirement for a good metric as we express it as a percentage and our unit of measure is detected attacks. Finally, we pass the "contextually specific" test since our metric measures security products in the context of other security products rather than just in isolation, which is a clear increase in value compared to existing metrics. This directly measures security product effectiveness in context allowing for an organization to base deployment decisions on our metric.

7 Future Work

Our proposed metric suggests many future areas of research to explore. We want to perform experiments to measure the benefit of using this metric compared to existing test frameworks that rely on testing layers of security separately. Also, exploring defense in depth strategies that maximize this metric seems promising. We hope to further investigate how to measure the value of redundancy in regards to diverse layers to increase the cost to attackers of evasion techniques. Working with existing organizations to implement our metric in their tests could greatly improve the accuracy of the value attached to certain security products being widely deployed today.

In the future, we also hope to expand our framework by adding attack cost information. For example, in [2] the authors expand the traditional metric of accuracy with added cost information such as attack severity. They also suggest varying responses rather than just naïvely blocking any suspicious data detected. Similarly, we could recalculate our metric with each attack weighted based on severity and the system's response. This approach would take the first steps towards a crucial translation of security measurements into actual organizational costs as certain severe attacks could be linked to financial losses. We want to explore how our metric coupled with detailed cost information could be translated into an economic measurement providing real return on investment estimates for security purchases.

8 Conclusions

Our proposed metric updates the tried and true metrics of detection rate and false positive rate with the context of multiple layers of security products. Our metric provides a more realistic picture than existing metrics by allowing individual security products to be measured in the context of other existing layers of security, which real environments are using. By comparing the ALDR of the set of security products with and without the security product to be evaluated we can measure the contribution of a single security product in its context such as with the rest of an organization's security products. A security product's detection rate and false positive rate are important, but more important is whether the security product is detecting attacks not easily detected by already deployed security products. This is the key value our metric provides the community.

References

[1] A. Jaquith. *Security Metrics: Replacing Fear, Uncertainty, and Doubt*. Addison-Wesley Professional, 2007.

[2] W. Lee, W. Fan, M. Miller, S. Stolfo, and E. Zadok. Toward cost-sensitive modeling for intrusion detection and response. *Journal of Computer Security*, 10(1/2):5–22, 2002.

[3] NSS Labs. http://www.nsslabs.com/.

[4] J. Pamula, S. Jajodia, P. Ammann, and V. Swarup. A weakest-adversary security metric for network configuration security analysis. In *Proceedings of the 2nd ACM workshop on Quality of protection*, QoP '06, pages 31–38, New York, NY, USA, 2006. ACM.

[5] A. Somayaji, Y. Li, H. Inoue, J. Fernandez, and R. Ford. Evaluating security products with clinical trials. In *Proceedings of the 2nd conference on Cyber security experimentation and test*, pages 3–3. USENIX Association, 2009.

[6] The Center for Internet Security. CIS Security Metrics 2010. http://www.cisecurity.org/metrics.html.