A Session Based Approach for Aggregating Network Traffic Data - The SANTA dataset

Charles Wheelus, Taghi M. Khoshgoftaar, Richard Zuech, Maryam M. Najafabadi
Florida Atlantic University
Email: {cwheelus, khoshgof, rzuech, mmousaarabna2013}@fau.edu

***Abstract*—This paper compares and contrasts the most widely used network security datasets; evaluating their efficacy in providing a benchmark for intrusion and anomaly detection systems. The antiquated nature of some of the most widely used datasets along with their inadequacies is examined and used as a basis for discussion of a new approach to analyzing network traffic data. Live network traffic is collected that consists of real normal traffic and both real and penetration testing attack data. Attack data is then inspected and labeled by means of manual analysis. While network attacks and anomaly features vary widely, they share some commonalities that are examined here. Among these are: self-similarity convergence, periodicity, and repetition. Further, the knowledge inherent in the definition of network boundaries and advertised services can provide crucial context that allows the network analyst to consider self-aware attributes when examining network traffic sessions. To these ends the Session Aggregation for Network Traffic Analysis (SANTA) dataset is proposed. The motivation and the methodology of collection, aggregation and evaluation of the raw data are presented, as well as the conceptualization of the SANTA attributes and advantages provided by this approach.**

I. Introduction

In just a couple of decades the Internet has transitioned from an experimental platform for interconnecting computers worldwide to an indispensable part of everyday life for a large portion of the global community. Banking, shopping, and all manner of essential services are now available online. As these Internet services have become ubiquitous in modern life, the illicit subversion of these services has become increasingly problematic. Our deep dependence on these services underscores the need for both the reliability and security of these network services. The detection of anomalous and malicious traffic is the cornerstone on which reliable online services depend upon.

Internet attacks and security breaches have become commonplace. According to the Cisco 2014 Annual Security Report [1], Security Alerts have continued to grow at a pace of 14% year over year since 2010, and have continued to grow in complexity as well. Unless anomalous and malicious traffic can be detected, it cannot be mitigated.

Interest in Internet security is nothing new and predates its commercialization in the mid 1990's. In 1997, Howard [2] provided an overview of Internet attack activity from the period of 1989 through 1996. His oft-cited work provides one of the earliest cohesive presentations of early Internet security incidents. Since then, the staggering growth of Internet traffic and the sheer volume of attack types, combined with the clear need for reliable Internet services has necessitated various and sundry methods for evaluating Internet traffic. In order to evaluate the effectiveness of different detection methods, datasets that contain different attack data are needed. While a few datasets have emerged over the years, they have proven to be sparse. Further these datasets suffer from many problems that make them less than ideal for the purpose of evaluating and verifying new and different attack detection methodologies. In addition, many of the most often used datasets are very old and largely irrelevant in the context of the attacks that besiege today's Internet services.

While machine learning techniques have enjoyed great success in areas such as product recommendation and spam detection, to name just a few, they have suffered from a misconception that, in terms of network security they are mostly effective in the limited context of detecting novel attacks. This perspective fails to consider the effectiveness of machine learning in recognizing similarity and patterns that may not be immediately clear from other methodologies. The motivation of this paper is to provide a different perspective by highlighting attributes of anomalous and malicious network traffic that set them apart from normal traffic, and this is done in the context of machine learning by including additional attributes (features) which are not available in other Intrusion Detection datasets which the authors are aware of. Commonalities often shared by anomalous and malicious traffic such as periodicity, convergence and repetition are explored. Attributes rooted in rudimentary network self-awareness are also considered, and attributes with self-awareness are those which emphasize network traffic of interest simply through their values or characteristics.

In prior datasets, a great emphasis was placed on data produced from penetration testing. While this approach is both valuable and necessary, it is contrived and fails to provide attack traffic that is realistic both in terms of variety and statistical distribution. Further, past efforts have also relied heavily on simulated normal data that suffers from

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.72

IEEE computer society

similar problems. The data and approach discussed herein places a large emphasis on real data, both attack and normal, that is lacking heretofore. Painstaking efforts to capture, analyze and manually label real world traffic data constitute the basis for the content of this paper.

The objective of this work is to produce a modern, next generation approach to evaluating network traffic for the purpose of recognizing network anomalous and malicious traffic patterns through the implementation of new metrics. To these ends the collection and analysis of real world Internet traffic is key in identifying attributes that can be used to benchmark network security measures. The SANTA dataset is proposed for this purpose; by means of the new attributes introduced in the SANTA dataset. The idea that anomalous and malicious traffic are fundamentally different from normal traffic, and share commonalties that can be measured and enumerated, is the rationale used to determine new attributes that might be likely to produce better results for data analytics. Many attacks, by their nature, are exceedingly repetitive and exhibit behavior different from normal traffic. Numerous attack patterns are analyzed and evaluated in order to determine what manner of attributes might result in the superior classification of such attack types.

The remainder of this paper is organized as follows. Section II presents related work and additional background regarding the objectives of this research effort as well as the rationale used to generate this new dataset. Section III presents the subject network and the methodology for collecting and analyzing the Internet traffic considered herein. Section IV presents the details of the SANTA dataset production, including the methodology used for the construction of sessions in IV-A and attribute definition in Section IV-B. Section IV-C explains the advantages of the approach used to produce the SANTA dataset and section IV-D discusses the labeled traffic patterns in the SANTA dataset. Discussion and future work are presented in Section V. Finally, conclusions are given in section VI.

II. Related Work

A common criticism of recent research efforts that seek to distinguish attacks from normal Internet traffic is that many employ the use of older datasets such as KDD CUP 99 [3] and NSL-KDD [4] for the evaluation of intrusion detection methods. In 1999 Hettich et al. [3], released the KDD CUP 99 dataset. KDD CUP 99 combines typical traffic features such as the port accessed, flags, the bytes to and from the source and destination, etc. as well as domain knowledge such as failed login attempts and attempts to gain administrative privileges. NSL-KDD, which was released in 2009 by Tavallaee et al. [4], used the same data as KDD CUP 99 but eliminates duplicate records in an attempt to provide a less biased statistical distribution. While these datasets may have proved to be useful over the years, they have grown increasingly irrelevant as new attack types and methods have emerged that are not reflected by these older datasets. McHugh [5] notes several problematic aspects of KDD CUP 99 including the appropriateness of evaluation techniques and a lack of details regarding the validation of the test data.

In recent years, datasets have emerged in an attempt to provide a better benchmark for use in intrusion detection. Unfortunately some of these datasets only utilize data created by means of penetration testing. Penetration testing is the act of exploiting a specific vulnerability on a predetermined portion of the network by generating the attack within parameters established by the penetration tester. This process involves identifying the vulnerability to be tested, determining how best to perpetrate the attack, execution, and then measuring the results. These penetration tests often provide actionable insight into the subject vulnerability being tested. While penetration testing is invaluable for producing desired attack vectors not readily available in the wild; non-simulated data is preferable whenever it is available. Mahoney and Chan [6] find that simulated data can lead to several problems, including that attributes that appear to be predictable in simulated traffic prove less than predictable in real traffic; leading to a greater probability of false positives. They also find that legitimate attacks are often masked in real traffic.

One notable and more recent dataset called the Kyoto dataset was built by Song et al. [7] with data from 2006 through 2009. The Kyoto dataset was built using various honeypot systems to collect known attack data and normal traffic was simulated using email and DNS servers. The Kyoto dataset used 14 of the same attributes as KDD CUP 99, as well as the addition of 10 new attributes. Kyoto does manage to overcome one problem of KDD CUP 99 by providing real attack data. Unfortunately Kyoto suffers from the fact that normal data is manufactured in an unrealistic manner that makes it a poor backdrop for the attack data provided in the Kyoto dataset. Kyoto normal data is comprised of traffic from a single server with a single domain and contains only email and DNS traffic. This data is too limited to provide a realistic variety or volume of normal traffic in comparison to the variety of attack traffic in the Kyoto dataset.

In 2012 Shiravi et al. [8], from the Information Security Center of Excellence, released the ISCX2012 dataset. The ISCX2012 dataset is a packet capture of both simulated attack data (based upon penetration testing) as well as simulated normal data based upon user traffic patterns and distributions. The packets are captured in a format that allows for later replay of the traffic. This approach sets ISCX2012 apart from previous datasets since it allows any user of the ISCX dataset performing repeated tests to use the same traffic each time. This ability to consistently repeat experiments is useful, since it allows for the traffic pattern to be repeated against varying intrusion detection methods and network topologies. The ISCX2012 dataset does not provide any new type of traffic attributes and the data itself does not provide session-based data; leaving this task of building session-based data to users of the ISCX dataset. Further ISCX2012 suffers from only

using penetration testing attack data, with no in the wild attack data.

Additionally, the unidirectional nature of ISCX is problematic since the context is buried in the payload data. In this dataset the matching of data into session based records for analysis is possible, but is left to the researcher as an exercise. Like communication between people, context is largely dependent on hearing both sides of the conversation. Network traffic is similar in the sense that the inbound and outbound network traffic typically results from one another. By analyzing the inbound and outbound flows separately, context can be obscured. While the underlying data is available in ISCX to produce session based data for analysis, there is no clear direction to consider the bi-directional nature of network traffic in the context of defined sessions; much less attributes that provide for this sort of analysis.

In 2010, Sommer and Paxson [9] proposed that network security has not enjoyed the same benefits from machine-learning techniques as other areas, such as product recommendation, optical character recognition, natural language processing and spam detection. They proposed that this is due to the faulty premise that machine learning analytic techniques are best suited for detecting novel attacks. They asserted that machine-learning techniques are better suited for recognizing the similarity between an already seen attack and a new attack without the need to precisely describe the new attack prior to its discovery. This point is extremely salient. The concept that anomalous and malicious traffic exhibits behavior that differs significantly from normal traffic is the mindset through which the attributes of the SANTA dataset were conceptualized and produced. In the course of identifying and evaluating various traffic patterns, this concept is observed repeatedly and is the basis for consideration of the new attributes of the SANTA dataset that are presented herein.

The motivation of this paper is to provide a new perspective on the collection and analysis of security traffic, as well as to produce a modern dataset for use in evaluating network intrusion and anomaly detection systems. In aggregating packet data and using flow data to facilitate the definition of sessions, the packets can each be classified as belonging to a particular session. From this point, metrics can be calculated based upon the packet and flow data, thus producing new attributes, such as periodicity, convergence, and repetition. Further by examining this data from the knowledge about the network, additional self-aware attributes can be enumerated as well.

While attack data generated from penetration testing can be uniquely valuable in producing data for specific attacks, reliance on only this type of data is clearly problematic since it is contrived and not necessarily reflective of an actual network scenario that might be encountered in the wild. Rather than relying only on penetration testing generated data, the SANTA dataset has been painstakingly inspected manually to identify actual attacks that occurred on the subject network. The abundant availability of attack data from the Internet is fortuitous in this sense.

III. Data Collection Methodology

The evaluation network used to collect the network data is a live operational network of a commercial Internet Service Provider (the "ISP"), with a mixture of varying types of Internet traffic. Some of the traffic is generated from customer networks that access the Internet through the ISP network, including email, browser, DNS, and all other types of Internet traffic that an average business might generate in the course of day-to-day operations. The ISP does not provide service to any consumer accounts. The ISP network also hosts a wide range of server traffic that is accessed by users from across the Internet. This includes hosted websites, email servers, DNS servers, and other various Internet services that are common to Internet Service Providers. The ISP is a host for businesses in dozens of countries around the world and has been in business for over two decades, which means there are a considerable variety of legacy software systems in operation on the ISP network. This and the global nature of the ISP, makes it an excellent candidate for the study of Internet based traffic. The operational nature of the ISP Network allows for the observation of in the wild network anomalies and attacks that originate from external unknown third parties. Additionally penetration testing is performed using modern attacks to specifically provide for newer vulnerabilities that do not exist in any of the prior security datasets mentioned herein.

It is worth noting that some of the traffic reviewed and considered as an attack could be the result of misconfiguration of services by a third party on the Internet and not the result of nefarious behavior; however from the perspective of the ISP, any traffic usurping resources and behaving outside of normal and expected parameters is sub optimal and costs unnecessary resources, resulting in a higher financial burden to the ISP. From this perspective, regardless of intent, all anomalies are therefore considered attacks.

The general topology of the ISP network (Figure 1) has two border routers that are connected to upstream network carriers. These border routers announce all of the IP addresses for the ISP network to the greater Internet. The ISP network is an autonomous system and employs Border Gateway Protocol ("BGP") to announce the autonomous system number ("ASN") and its corresponding IP address space. These two routers serve as the demarcation point for differentiation between packets that originate from an external source (an "external address") and packets that originate from an internal source (an "internal address"). Outbound packets are defined as those that originate from any IP addresses announced by the ISP ASN. Inbound packets are all other packets. This clear border defined by tracking the difference between inbound and outbound packets, as well as actively advertised ports and addresses vs. dark addresses constitutes, in a basic sense, a primitive but important form of self-awareness exhibited within the ISP network for the purpose of examining different types of anomalous network traffic.

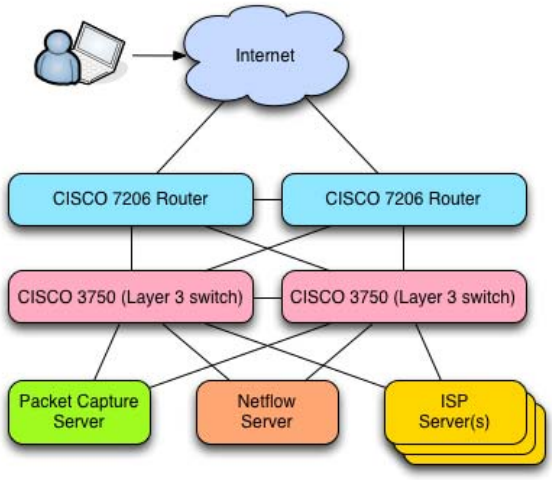

Figure 1 ISP Network Topology

While dark IP addresses are part of the internal IP address space that is announced using BGP, they remain unassigned to any customer or ISP network device and not advertised to external IP addresses as having any active services available. They are interspersed with active addresses and reveal targets that are clearly only receiving attack traffic. Traffic that targets them inadvertently provides insight into the intent of the source traffic. Source IP addresses that attempt to interact with dark addresses can only be anomalous or malicious in nature; in either case it is reasonable to conclude that traffic from the source IP is suspect and a good candidate for closer inspection. This technique is simple yet powerful since it makes it difficult for an attacker to engage in subterfuge from the same IP address after scanning addresses that are dark, and clearly illustrates how a feature like dark IP addresses can promote self-awareness of suspect behavior. This concept is a direct result of the fact that we are only considering traffic that crosses the border of the ISPs Network, and no internal traffic (from one internal host to another internal host) is being included in the dataset.

In this context, self-awareness is attributable to the clearly defined set of IP addresses, as well as their role in the functioning of the ISP network. IP addresses announced by the border routers are internal; all others are external. IP addresses announced via BGP that are known to not advertise any service to external IP addresses, and remain unassigned to ISP customers, are dark addresses that should not be the recipient of traffic from external IP addresses.

Certain internal IP addresses are known to provide services on pre-defined ports. For example: DNS is known to run on port 53 of certain known internal IP addresses. Traffic received on ports other than 53 on these internal IP addresses is identifiable as scan traffic since no other services are running on those IP addresses. These properties of, and relationships between, IP addresses allows for clear context in the mind of an analyst reviewing the traffic data.

All of the external facing router ports on the ISP network are aggregated into a single stream for data capture. Two types of raw data are collected: packet capture and NetFlow data. The raw packet capture is literally a copy of each network frame that traverses the demarcation point. This bit for bit copy of the network stream allows for deep analysis of each network packet, including the payload and all of the packet header data required to transmit the packet to the intended destination. Given that fact that we are only concerned with traffic entering and leaving the ISPs Network from the Internet, only data aggregated when traversing the externally facing network ports is considered

Given that the data payload is available in packet capture, many intrusion and anomaly detection methods employ a method known as deep packet inspection to determine if the traffic of which the packet is a component is likely to be malicious in nature. While this approach has its benefits, it is computationally expensive, as discussed by Porter [10], and may present scalability issues for larger networks. Furthermore, in the event that the payload is encrypted, the problem becomes intransigent, as the intent of the payload may not be discernable in any reasonable timeframe, or even at all. This points to a clear need for the ability to assess the nature of the traffic based solely upon other means such as the packet header, the NetFlow data, or data derived therefrom.

NetFlow was originally introduced by Cisco Systems [11] in 1990, for the purpose of network traffic planning and growth management. While packet capture gives the most granular perspective of network traffic, NetFlow provides a higher level of abstraction for grouping related packets together into groups commonly referred to as flows. Generally speaking a flow is a group of packets that all share the same source and destination IP address, as well as the same source and destination port, and also the same protocol. NetFlow data is commonly stored in a binary format with the ability to display the flows in different aggregation schemes based upon the desired perspective. This gives the researcher the ability to produce different views of the flow data based upon the aggregation syntax used. This ability makes NetFlow very powerful and useful for intrusion and anomaly detection. A good example of the difference between captured packet inspection and NetFlow would be viewing a forest by hiking through the forest as opposed to flying over the forest in a hot air balloon. Each perspective provides details not apparent from the other.

Collection of NetFlow employs Cisco NetFlow version 9 from the ISP Network routers, which are then aggregated using the NFDUMP [12] NetFlow collector. The records are later dumped into text format using NFDUMP. For packet collection a mirror port for each outbound router port is aggregated and collected using tshark [13] in pcapng [14] format. Wireshark [15] is used for the manual inspection of the captured packet data. Due to the scale of the project, and time and resource constraints, only data that traverses in ISP network boundary is considered. For attacks generated using

penetration testing techniques, a combination of custom tools and metasploit [16] are used.

IV. THE SANTA DATASET

The SANTA dataset marks a different approach to developing a network security dataset for analysis by intrusion anomaly detection algorithms. It is obvious that attacks can originate from any source, internal or external. These sources can vary from attacks by network-based botnets, internal workstations that are infected by malware, trusted network users with malicious intent, or a wide range of other sources. While the intent of the SANTA dataset design is for use with the external network attack as the focus, the approach outlined herein could be easily translated to a different context for other scenarios and topologies by simply adding more NetFlow and packet capture points inside the network to be studied.

The problem with both NetFlow and packet data is that, in their raw state, they do not provide the perspective of the entire session of network communications. The unidirectional concept of packets and flows provide incomplete context for the traffic of which they are constituents. The focus of the SANTA dataset is to provide both greater context by combining the use of NetFlow data, packet data, and known network constraints, such as dark IP addresses.

A. Network Sessions

Sessions are constructed by grouping packets together based upon the NetFlow data. Each inbound flow is matched with a corresponding outbound; all packets that contain all of the associated attributes of that flow data are categorized as part of that session. This means, for example, that an inbound DNS request packet produces a flow record in the NetFlow data that is later paired with the corresponding outbound flow record. Once these flow records are paired, they produce a session. Once grouped by session, all further metrics can be attributed to the packets that comprise the session to which the packet belongs.

The session is defined as beginning at the earlier NetFlow timestamp of two corresponding flow records and ending after adding the duration of the later corresponding flow record. The flow records must also match by protocol; meaning flow records that contained the same IP address pairs, but different protocols would be separate sessions. All packets within the defined time range that meet these criteria are considered part of the same session.

The genesis of additional session features resulted from considerable time spent reviewing captured packets and NetFlow data. By using various filtering techniques, groups of packets and flows were isolated, studied and categorized. During this process, it became evident that certain attack types had characteristics that caused them to stand out from other traffic. For example, after filtering in order to only display DNS traffic, and then filtering on various IP addresses that are assigned to the DNS servers of the ISP Network, a DNS amplification attack became immediately obvious.

A DNS Amplification attack forges the source IP address (known as spoofing) so that the resulting DNS result packet is sent to the target of the attack. In this instance the resources of the intermediate host are used to attack the spoofed address. The inbound packets to the intermediate host are considerably smaller than resulting outbound packets that are sent to the attack target, thus the attack is “amplified”.

The repetitive and periodic nature of the DNS amplification attack, which is required for the attack to be effective, allows for the quantification of new attributes that provide metrics for repetition and periodicity. In other words, these attacks, by their very nature, will provide very different values for attributes that measure repetition and periodicity than will normal traffic. Further the packets of this session exhibit a very high degree of self-similarity, thus allowing for self-similarity measurements elucidated in terms of convergence. DNS amplification also differs from normal DNS traffic in terms of velocity. This can be measured in terms of packets per second, bits per second and bytes per packet. Such context makes the anomalous nature of the DNS amplification abundantly clear without inspecting payload data. In hand labeling this attack as DNS amplification, the packet payloads were inspected to confirm the attack is indeed a DNS amplification attack. While packet payloads were used in verifying the hand labels for the sessions, they are not included in the final dataset.

Similar methods were used in evaluating other attack patterns. For example, the penetration test for the Heartbleed attack described later produced data that while encrypted, and therefore not suitable for deep packet inspection, produced patterns that should be recognizable based upon evaluation of the SANTA attributes. This example highlights a very serious problem with relying on payload data for traffic analysis. Encrypted traffic is very common and is not in and of itself cause for suspicion. It is therefore a viable technique for an attacker to obscure the intent of his traffic by encrypting the payload data.

B. Attribute Definitions

Due to the fact that the SANTA dataset attributes are describing sessions rather than flows or packets, consideration was given to how best to produce the most descriptive and useful attributes. While a few of the attributes mentioned later, in Table 1, are analogous to prior datasets, most of the SANTA attributes are unique. Sessions are identified and described in terms internal and external IP address, internal and external target port and protocol. A Boolean variable named IO match confirms that each inbound flow corresponds to an outbound flow. While it may seem counterintuitive that a session would not have matching inbound and outbound flows, The RUDY attack described in section IV actually exhibits this property, and yet it is legitimate traffic that complies with the HTTP protocol.

Table 1 Attributes of the SANTA Dataset

Attribute	Description
External IP	The external IP address (target or source)
Internal IP	The internal IP address (target or source)
Internal target port	The port number of the internal target
External target port	The port number of the external target
Attack label	The type of attack if known (detailed in "Labeled Traffic Patterns" below). Any traffic that is not from a known attack type is labeled as unknown.
Protocol	Transmission protocol
IO match	Whether the inbound flow has an associated outbound flow record (Boolean)
Duration	The elapsed time, from the earliest of the associated inbound or outbound flow until the end of the later flow
Session size in bytes	The total size for the session in bytes
Session size in packets	The total size for the session in packets
Inbound session convergence	Self-similarity of the inbound packets in the session is determined by examining the variance in size of the inbound packets
Outbound session convergence	Self-similarity of the outbound packets in the session is determined by examining the variance in size of the outbound packets
Session repetition	The packet count, per session, of all packets that are of the most common packet size
Session periodicity	The measure of periodicity within a session (given by the variance of timestamp differences between inbound packets)
Inbound velocity pps	Velocity of inbound traffic measured in packets per second
Inbound velocity bps	Velocity of inbound traffic measured in bits per second
Inbound velocity Bpp	Velocity of inbound traffic measured in bytes per packet
Outbound velocity pps	Velocity of outbound traffic measured in packets per second
Outbound velocity bps	Velocity of outbound traffic measured in bits per second
Outbound velocity Bpp	Velocity of outbound traffic measured in bytes per packet
RIOT packets	Ratio of inbound to outbound traffic measured in packets
RIOT Bytes	Ratio of inbound to outbound traffic measured in bytes
Dark access	The inbound target IP address is dark (Boolean)
Port scanner	The session external source address has previously attempted to access an inactive port (Boolean)
Dark scanner	The session external source address has previously attempted to access an IP address that is dark (Boolean)
Originating ASN	The ASN that announced the external source IP address.
ASN scanner	One or more IP addresses from the originating ASN are marked as exhibiting scanning behavior
Flags	All TCP Flags used in the session

While prior datasets described the data in terms of source and destination size, the SANTA dataset combines these into session size, and instead includes the ratio of inbound to outbound traffic, described both in terms of traffic size in bytes as well as the number of packets.

The SANTA dataset provides particularly unique attributes to describe both the self-similarity of inbound packets comprising a session in terms of size, described as inbound session convergence; and the outbound session convergence which described self-similarity of the outbound packets.

While prior datasets do contain a rudimentary measure of velocity, the SANTA dataset provides far greater granularity by defining three different measurements of both inbound and outbound traffic velocity; packets per second, bits per second and bytes per packet. This provides for three concurrent perspectives on traffic velocity for all records in the SANTA dataset.

The SANTA dataset was designed to take advantage of basic self-awareness by providing attributes based not only upon the knowledge of the network boundary in terms of IP addresses, and also the basic usage of the internal IP addresses. This basic self-awareness allows for the creation of network attributes such as dark access, dark scanner and ASN scanner described below.

It should be noted that rather than considering source and destination concepts of network traffic, the SANTA dataset instead is concerned with internal and external IP addresses. In other words, the approach in constructing the SANTA dataset is to view all traffic in the context of its origin, and whether or not the originator is internal or external. The fact that traffic originates from an external source provides contexts as to its possible nature. For example, a DNS request that originates from an external source indicates that an entity outside of the internal network seeks DNS information about a service presumably from the internal network; whereas a DNS request that originates from an internal address, targeting an external DNS server seeks information about services provided by external resources. The SANTA dataset views all addresses in light of the question: Is this address part of my network, or is it not?

Similarly, the target port is generally speaking the only port worth serious consideration due to the fact the ports used for return traffic are generally selected in an arbitrary fashion while target ports generally follow a convention in terms of the type of service offered and can therefore provide context to the session. Of the internal and external IP address in the session, it is the recipient of the traffic to a standard port is the target.

For example, the observation of a periodic series of packets to the inbound target port of 53, each of which is followed by a return packet, would logically be recognized as DNS queries and answers, the repetition, combined with the fact that DNS query results are typically cached by the originating computer; might indicate anomalous behavior. The higher the frequency and the more periodic this behavior is, the greater the probability that this traffic is malicious in nature. This is the sort of logic typically followed by security personnel when evaluating network security data. The intent of the SANTA dataset is to describe the network session data in a fashion that lends itself to successful network intrusion and anomaly detection analysis.

Both Kyoto and KDD CUP 99 contain an attribute known as duration that is analogous to the duration attribute in the SANTA dataset. The Internal Target Port and External Target Port are in essence the same as the Service attribute in KDD CUP 99 and Kyoto, except that they also describe the targets inclusion or exclusion of the ISPs Network. In other words, the presence of a pre-defined service on a standard port inherently defines the direction of the traffic as inbound or outbound. In the SANTA dataset, the attribute known as Session Size in Bytes is the sum of the src_bytes and dest_bytes, and the Source Bytes and Destination bytes attributes from KDD CUP 99 and Kyoto respectively. The

SANTA dataset also contains a protocol attribute as found in KDD CUP 99.

During the process of evaluating the underlying data, common characteristics emerged that would become some of the new attributes for the SANTA dataset. After extensive review of DNS amplification attacks, the differences between these attacks and normal DNS traffic was abundantly clear. Normal DNS is considerably aperiodic. When a DNS record is requested, the requesting server caches the result so that future requests do not require an additional lookup from the authoritative DNS server. This design reduces both the load on the authoritative server and the likelihood of network congestion. A DNS amplification attack can only succeed by repeatedly requesting a record that is redirected to a spoofed return address. The resulting periodicity and repetition makes the traffic definitively anomalous. In Figure 2, visual inspection of normal and attack traffic demonstrates immediate visual recognition that the attack traffic is abnormal. All red traffic is a DNS amplification attack with the same, spoofed return IP address. This traffic is highly convergent, meaning the size of each inbound DNS request packet is the exactly same and each outbound packet is also the same size. The pattern is repeated and is clearly, but not perfectly periodic. Each of the four green sessions represents a normal DNS query: a single query packet, and a single response, from a unique external IP address. Each of these sessions clearly exhibits variance in size for both inbound and outbound traffic and is not repeated.

These two types of traffic are in stark contrast to one another. Not only in terms of convergence and periodicity, but in terms of the difference between the ratio of inbound to outbound traffic for normal DNS traffic vs. the DNS amplification attack traffic. A typical, normal DNS query and response would usually return a response packet that is roughly 2 to 3 times as large in terms of packet size, and would not repeat itself in a short period of time. A DNS amplification attack would reoccur and would have a response packet that might be a factor of 8 or more as large. In one case, a DNS amplification attack was observed that repeated over five thousand times in a five-minute period. Again, upon visual inspection, in Figure 2 we see the response packets that are nearly an order of magnitude greater in terms of packet size.

Clearly different attack types will differ with respect to the SANTA dataset attributes, but the goal is to express attributes that clearly distinguish between attack or anomalous traffic and that of normal traffic. In another case, the RUDY (aRe yoU Dead Yet) attack (see section IV-D, below) provides an example that, while very different in nature, makes good use of the SANTA dataset attributes. RUDY takes advantage of the fact that an HTTP post operation allows for the connection to remain open indefinitely in cases where the post data arrives very slowly; for example one byte per second. While this design was an accommodation to allow slow dial-up Internet connections to post data, it is exploited by RUDY as a denial of service. This is an interesting attack and is somewhat difficult to detect since it uses a very small amount of bandwidth and complies perfectly with the HTTP protocol. Even so, the RUDY attack will still exhibit an exceedingly high ratio of inbound to outbound HTTP traffic, exhibit a very low Inbound velocity Bpp, and will be quite periodic and self-similar. This will make the RUDY attack a clear outlier in comparison to normal HTTP traffic. This is due to the fact that normal HTTP traffic is a series of requests and responses, while RUDY is one unending request that can never receive a response.

These two examples highlight the general approach in developing the SANTA dataset attributes. The concept is to differentiate the behavior inherent in normal traffic from anomalous and attack traffic in a manner that produces attributes that exhibit clear numerical differences. This approach differs significantly from the prior datasets mentioned herein.

One factor common to both the RUDY and DNS amplification attack is the repetitive and periodic nature of the traffic. In observing these attacks, the periodic behavior of the repeating patterns was striking, especially given that the attacks are very different in nature. While neither displayed what might be considered perfect periodicity, the pattern was still visually obvious. Three possible approaches are proposed for measuring the attribute of periodicity. The simplest approach would be to measure the variance of the difference in time intervals between each inbound self-similar packet that belongs in the same session. This variance of the difference between timestamps is used as the periodicity attribute.

Another approach for calculating periodicity would be to perform a Fast Fourier Transform (FFT) on the frequency of inbound packets. Barbosa et al. [17] proposed using FFT for detecting attacks and anomalies on Supervisory Control and Data Acquisition (SCADA) networks, and they found that attack traffic disrupts the periodicity of SCADA network traffic. In comparison to SCADA networks, Internet traffic is relatively random; this implies the likelihood that the periodic nature of some attacks will differ significantly when compared with the majority of Internet traffic.

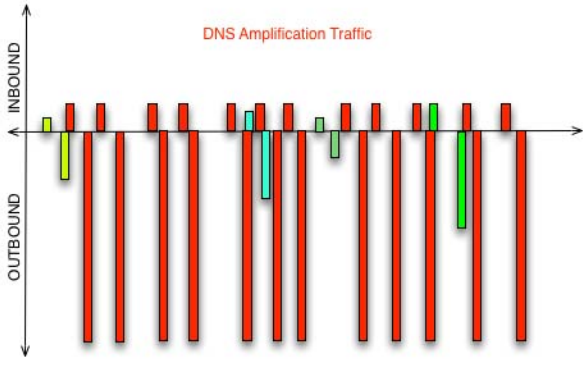

Figure 2 DNS Traffic

Berberidis et al. [18] describe the challenges of detecting periodic behavior in time series analysis. They propose the utilization of autocorrelation using FFT as a filter. Their approach has the benefit of not requiring any prior knowledge of the data in determining all of the periodicities of data in a time series. While their experiment consisted of analyzing supermarket and power consumption data, the same approach could also be used for evaluating the periodicity of network traffic.

C. Advantages and Limitations

All of the traffic in the SANTA dataset is generated from actual Internet traffic. While some of the attack data was generated by attacks made from external hosts controlled by the author, actual attacks were conducted against internal hosts. No traffic is assumed to be normal. Traffic flows not specifically labeled as an attack type are labeled as unknown. The attributes in the SANTA dataset are generated from elements of both packet capture and NetFlow data, combining the strengths of both. The sessions are defined using NetFlow standards and are not arbitrarily determined. All labels are manually assigned based upon evaluation and manual confirmation of attack type. As an example, the DNS amplification attack is obvious due to the frequency, repetition and self-similarity of the session as well as the ratio of the inbound and outbound traffic volume. In addition to these factors, the payloads are inspected to confirm that the attack is actually a DNS amplification attack.

The topology used to capture packets and NetFlow data only utilize data that traverses the border routers. Due to this limitation, no internal traffic data is captured. As mentioned previously no payload data is included. While the payload data is present in the original packet capture, the SANTA dataset is a hybrid representation that uses elements of both packet capture and NetFlow. Due to this fact, payload data, while used for the initial verification of the hand labels for some of the attacks, is not included in the final session data. Admittedly, the review of packet and NetFlow data is time intensive and the production of the resulting data for SANTA is resource intensive. Real-time approaches to threat mitigation may not be practical with this approach, unless sufficiently modified. However, experimentation with some of these features could possibly yield real-time benefits through more research and refining the methodologies. Additionally it is conceivable that manual labeling may introduce noise by incorrectly classifying traffic. Even though every effort is made to verify the accuracy of labeling, humans are prone to error.

D. Labeled Traffic Patterns

A DNS Amplification attack, as discussed by Kambourakis et al. [19] forges the source IP address (known as spoofing) so that the resulting DNS result packet is sent to the target of the attack. In this instance the resources of the intermediate host are used to attack the spoofed address. The inbound packets to the intermediate host are considerably smaller than resulting outbound packets that are send to the attack target, thus the attack is "amplified". Kambourakis et al. examine methods of detecting these attacks. This attack, while not necessarily new, is not included in any of the prior datasets mentioned herein.

Table 2 Advantages of SANTA Dataset

Advantages	SANTA	KDD	Kyoto	ISCX
Realistic normal traffic (not simulated)	YES	NO	NO	NO
Penetration testing attack traffic	YES	YES	NO	YES
Real, in the wild attack traffic	YES	NO	YES	NO
Modern attacks	YES	NO	NO	NO
Manually inspected and verified attack labels	YES	NO	NO	NO
Periodicity attributes	YES	NO	NO	NO
Repetition attributes	YES	NO	NO	NO
Convergence attributes	YES	NO	NO	NO
Velocity attributes	YES	YES	YES	NO†
Self-aware scanning attributes	YES	NO	NO	NO

† Implied (Can be calculated)

One particularly problematic layer 7 DDOS attack known as RUDY (aRe yoU Dead Yet) studied by Damon et al. [20] exploits a weakness in the HTTP protocol. The attacker begins an HTTP POST operation and then delays the completion of the transaction by sending very small packets to keep the session open without completing the POST operation. This vulnerability exists to accommodate the slower dialup connections that are mostly non-existent today. Damon, et al. note the relatively modest amount of resources required to effectively execute this attack. This attack has existed for quite some time but is not included in any of the prior datasets mentioned herein. The subtle nature of the attack makes it an interesting case for study and was the primary motivation for its inclusion.

Haggerty et al. [21] reviewed and discussed the SYN flood attack; a denial of service attack that attempts to usurp resources by sending repeated SYN TCP packets to the destination host while ignoring the resulting ACK TCP packets, and never sending the SYN ACK packets that would complete the establishment of a normal TCP session. The resulting broken TCP sessions continue to take up network sockets until they time out. While this is also a relatively simple attack, it remains common as well so it is included in the SANTA dataset; and is included in all of the datasets mentioned herein.

An ICMP flood as discussed by Limwiwatkul et al. [22], also known as a ping flood is a denial of service attack where an unusually large number of ICMP packets are sent to the destination network with the intent of consuming network resources to prevent the timely functioning of the network. The ICMP flood is a simplistic attack, yet it is still common and so it is included in the SANTA dataset.

The Heartbleed attack, the implications and repercussion of which are discussed at length by Kamp [23] uses a flaw discovered in several versions of the openssl library. This weakness was revealed publicly in April 2014 and was introduced by a bug in 2011. The vulnerability allows the

attacker to read the contents of data that is supposed to be protected by encryption. Due to the popularity of the openssl, the vulnerability was widespread. This attack is very new and is not included in any of the datasets mentioned herein.

A Port Scan as discussed by Sridharan [24] is not really an attack per se, but merely a probe of available services on a host. While this is not an attack in the strictest sense, it is often categorized as such since it is often a precursor to an attack. In this scenario the attacker seeks to determine what services may be attacked based upon responses from connection attempts to various ports on the host. Similarly, a Dark Scan, studied by Sellke et al. [25], is an attempt to connect to an IP address that does not have any publicly advertised services running.

In the SANTA dataset, all traffic not specifically recognized as one of the aforementioned attack types is labeled as Unknown. These attack types were chosen based upon current frequent attack types on the ISP network and attack types that are newer or difficult to spot using conventional means. The RUDY and Heartbleed attacks were produced by means of penetrations testing and all others were discovered in the dataset and were labeled by manual observation

V. Discussion and Future Work

The attributes of the SANTA dataset were chosen in an attempt to provide new metrics that are likely to provide for greater success in training intrusion and anomaly detection algorithms to predict attack types. Earlier datasets typically suffer from poor normal or attack traffic. Most are somewhat antiquated and lack one or more of the following attribute types: velocity, periodicity, convergence, repetition or self-awareness. While all prior datasets do consider source and destination traffic size in some fashion, the importance of the ratio between the two is not elucidated or even implied. In contrast, the SANTA dataset provides the ratio of both the traffic size and packet count.

Velocity attributes for the SANTA dataset are defined in a far better fashion than in prior datasets by providing bits per second, packets per second, and bytes per packet for both inbound and outbound traffic. This greater level of granularity allows for an in depth view of traffic velocity. This attribute is an important feature as explained in the examples included herein.

Self-similarity is clearly an important attribute to consider as illustrated in the examples provided herein. Further, the inclusion of repetition and periodicity attributes provides an as of yet overlooked metric for consideration when evaluating network traffic. Additionally, the inclusion of self-aware attributes such as dark access, dark scan, port scan, and ASN scanner provide a new level of sophistication heretofore unavailable in prior datasets.

Convergence is a markedly important, yet heretofore unused, property of network traffic, especially when considering that several attack vectors such as Heartbleed, SYN flood, ICMP flood, and DNS amplification display highly convergent traffic patterns. These attacks also display highly repetitious and periodic traffic patterns, thus being clear outliers from normal traffic. Additionally, when considering the session size, the inclusion of the RIOT attributes provides context as to the difference between inbound and outbound traffic. This is especially true for attacks like RUDY that are overwhelmingly biased towards input traffic, yet have a low velocity in comparison with most other network traffic.

Clearly the SANTA dataset provides several distinct advantages not available in KDD, Kyoto, or ISCX. By defining sessions based upon NetFlow and packet data, the SANTA dataset provides greater context for the network traffic from which it was derived. The painstaking inspection, verification and hand labeling pay dividends in the ability to accurately produce new attributes that can be utilized to detect anomalies and intrusions. Specifically, the addition of the various velocity, periodicity, repetition, convergence, and self-awareness attributes provide greater context and therefore a greater probability of attack or anomaly detection.

The list of attack types is obviously far from exhaustive. Future work will include likely include new attack types and possibly new attributes and the production and analysis of various periodicity measurements. The performance of the SANTA dataset attributes introduced herein using various data analytic techniques is also a likely candidate for future work. Internal traffic may also be considered if feasible.

VI. Conclusions

This paper presents the results of a comprehensive network data collection effort from the network of a commercial Internet Service Provider of both NetFlow and packet capture data. Both in the wild attacks discovered in the traffic data as well data from penetration testing are analyzed using various tools and filtering techniques. The analysis of the attack data yields several commonalities amongst different attack types. Sessions are defined and analyzed; motivated by the desire to produce a more complete perspective of the nature of attacks. The nature of these commonalities is considered in the production of new network session attributes.

A comparison of these new attributes and attack types, in consideration of older attack datasets, is discussed; and there are several advantages from the approach taken in this paper. One benefit is the availability of real normal traffic data as opposed to simulated normal traffic data of prior datasets. Another advantage is the availability of real attack traffic data whenever possible as opposed to penetration testing traffic data of some prior datasets. Future research can benefit from the availability of new and current attack traffic data unavailable in prior datasets. The convergence, repetition and periodicity attributes of the SANTA dataset provide a new perspective to commonalties of network attack traffic. Machine Learning can take advantage of the greater granularity of velocity attributes from the SANTA dataset.

The availability of the self-aware scanning attributes of the SANTA dataset could be beneficial for security analytics, and may even be applicable to real-time applications.

The first main contribution of this research is to provide a methodology that may improve Intrusion Detection through new approaches and attributes. While manually inspecting attacks in a network session, it became apparent that certain attributes are useful for detecting certain attack vectors and hopefully these results can be generalized and applied usefully with future work. The second main contribution is that a new benchmark dataset is being introduced for future research with the motivation to overcome the shortcomings of mainstream Intrusion Detection datasets. Obviously, the authors do not contend that this dataset is perfect either, but it does bring unique insights that do not currently exist for Intrusion Detection datasets.

REFERENCES

[1] Cisco Systems '2014 Annual Security Report'. Cisco. Accessed May 2014. http://www.cisco.com/web/offers/lp/2014-annual-security-report/index.html

[2] Howard, John D. *An analysis of security incidents on the Internet 1989-1995*. CARNEGIE-MELLON UNIV PITTSBURGH PA, 1997.

[3] Hettich, Seth, and S. D. Bay. "The UCI KDD Archive [http://kdd. ics. uci. edu]. Irvine, CA: University of California." *Department of Information and Computer Science* (1999).

[4] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," *Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 2009.

[5] 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*, vol. 3, no. 4, pp. 262–294, 2000.

[6] Mahoney, Matthew V., and Philip K. Chan. "An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection." *Recent Advances in Intrusion Detection*. Springer Berlin Heidelberg, 2003.

[7] Song, Jungsuk, Hiroki Takakura, Yasuo Okabe, Masashi Eto, Daisuke Inoue, and Koji Nakao. "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation." In *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pp. 29-36. ACM, 2011.

[8] Shiravi, Ali, Hadi Shiravi, Mahbod Tavallaee, and Ali A. Ghorbani. "Toward developing a systematic approach to generate benchmark datasets for intrusion detection." *Computers & Security* 31, no. 3 (2012): 357-374.

[9] Sommer, Robin, and Vern Paxson. "Outside the closed world: On using machine learning for network intrusion detection." *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010.

[10] Porter, Thomas. "The perils of deep packet inspection." *Security Focus* (2005).

[11] Claise, B. "Rfc 3954: Cisco systems netflow services export version 9 (2004)." *Retreived online: http://www. ietf. org/rfc/rfc3954. txt* (2007).

[12] Haag, Peter. "Watch your Flows with NfSen and NFDUMP." *50th RIPE Meeting*. 2005.

[13] Nascimento, Gustavo, and Miguel Correia. "Anomaly-based intrusion detection in software as a service." *Dependable Systems and Networks Workshops (DSN-W), 2011 IEEE/IFIP 41st International Conference on*. IEEE, 2011.

[14] Kuga, Yohei, et al. "EtherPIPE: an Ethernet character device for network scripting." *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking*. ACM, 2013.

[15] Lamping, Ulf, and Ed Warnicke. "Wireshark User's Guide." *Interface* 4 (2004): 6.

[16] Foster, James C. "Metasploit Toolkit for Penetration Testing, Exploit Development, and Vulnerability Research." (2007).

[17] Barbosa, Rafael Ramos Regis, Ramin Sadre, and Aiko Pras. "Towards periodicity based anomaly detection in SCADA networks." IEEE Industrial Electronics Society, 2012.

[18] Berberidis, Christos, et al. "Multiple and partial periodicity mining in time series databases." *ECAI*. Vol. 2. 2002.

[19] Kambourakis, Georgios, et al. "Detecting DNS amplification attacks." *Critical Information Infrastructures Security*. Springer Berlin Heidelberg, 2008. 185-196.

[20] Damon, Evan, et al. "Hands-on denial of service lab exercises using SlowLoris and RUDY." *Proceedings of the 2012 Information Security Curriculum Development Conference*. ACM, 2012.

[21] Haggerty, John, et al. "DiDDeM: a system for early detection of TCP SYN flood attacks." *Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE*. Vol. 4. IEEE, 2004.

[22] Limwiwatkul, Lersak, and Arnon Rungsawang. "Distributed denial of service detection using TCP/IP header and traffic measurement analysis." *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*. Vol. 1. IEEE, 2004.

[23] Kamp, Poul-Henning. "Quality Software Costs Money-Heartbleed Was Free." *Queue* 12.6 (2014): 10.

[24] Sridharan, Avinash, Tao Ye, and Supratik Bhattacharyya. "Connectionless port scan detection on the backbone." *Performance, Computing, and Communications Conference, 2006. IPCCC 2006. 25th IEEE International*. IEEE, 2006.

[25] Sellke, Sarah H., Ness B. Shroff, and Saurabh Bagchi. "Modeling and automated containment of worms." *Dependable and Secure Computing, IEEE Transactions on* 5.2 (2008): 71-86.