Proceedings of the 2005 IEEE
Workshop on Information Assurance and Security
United States Military Academy, West Point, NY

Flow Based Observations from NETI@home and Honeynet Data

Julian B. Grizzard, Charles R. Simpson, Jr., Sven Krasser, Henry L. Owen, George F. Riley
{grizzard, rsimpson, sven, owen, riley}@ece.gatech.edu
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250, USA

Abstract—
We conduct a flow based comparison of honeynet traffic, representing malicious traffic, and NETI@home traffic, representing typical end user traffic. We present a cumulative distribution function of the number of packets for a TCP flow and learn that a large portion of these flows in both datasets are failed and potentially malicious connection attempts. Next, we look at a histogram of TCP port activity over large time scales to gain insight into port scanning and worm activity. One key observation is that new worms can linger on for more than a year after the initial release date. Finally, we look at activity relative to the IP address space and observe that the sources of malicious traffic are spread across the allocated range.

I. Introduction

The Internet has grown from the small ARPANET to an unfathomably large network. As with any new technology, the Internet has grown from its infancy to a stage where security concerns become a considerable problem. Today's Internet is plagued with a plethora of worms, viruses, malware, spam, and otherwise malicious traffic. In this paper, we make observations about end user Internet activity by comparing honeynet traffic and NETI@home traffic in order to better understand the security problems of the Internet.

Our strategy for understanding the malicious Internet traffic is a flow based analysis of several years of honeynet data and NETI@home data. We study a number of metrics visually over large timescales and plot both the honeynet dataset and the NETI@home dataset and then compare the results. Some interesting points include flow activity across the IP address space, port scan activity, new and lingering worm traffic, as well as other observations. Below we provide some background information on the datasets used.

A. NETI@home

The NETI@home project was started to collect end user statistics from hosts on the Internet. These measurements are gathered using an open-source software package that end users can download from the NETI@home website [1]. The software package has been designed to run on a number of platforms in order to reach as many different users as possible. To collect data, Internet users must volunteer to run the software package on their end hosts. Once the package is installed, the NETI@home client will collect network statistics from the end host and periodically send a report back to the NETI@home server.

The NETI@home project collects statistics on the TCP, UDP, ICMP, and IGMP protocols. Users can select a privacy level of high, medium, or low, which determines what portions, if any, of the IP addresses are recorded in each flow. Some of the analysis presented in this paper requires using only low or medium privacy statistics and may skew the results slightly, but we feel that our user base is large enough that such skewing is minimal.

The NETI@home dataset we are analyzing was collected from June 1, 2004 to February 28, 2005 and consists of reports from at least 500 uniquely identifiable users. There are approximately 31 million TCP flows and 33 million UDP flows in this dataset, constituting 65 gigabytes of transferred network traffic. The remaining flows consist of 600 thousand ICMP flows and 250 thousand IGMP flows.

B. Georgia Tech Honeynet

A honeynet is a network of resources whose value lies in the illicit use of those resources. All network traffic to and from a honeynet is suspicious, but a small amount of traffic may be legitimate. However, most of the traffic on a honeynet is malicious in nature.

The Georgia Tech Honeynet Project was launched in the summer of 2002 and immediately began collecting data [2]. The dataset we are using consists of nearly three years of honeynet traffic with very few service interruption points for maintenance and upgrades. All network traffic to and from the honeynet has been logged and archived, including the traffic between the honeypots.

To better understand the conclusions we draw from this data, it is important to understand the network on which this honeynet has been deployed. There are over 15,000 students enrolled at Georgia Tech and approximately 5,000

0-7803-9290-6/05/$20.00 ©2005 IEEE.

staff and faculty employed. The supporting network consists of more than 40,000 networked systems all within Georgia Tech's ".edu" address space. The honeynet has been deployed within this ".edu" address space and is accessible from internal machines within the Georgia Tech address range as well as external machines.

The honeynet dataset we are analyzing was collected from August 19, 2002 to February 28, 2005 and consists of reports from 38 unique IP addresses. There are approximately 2 million TCP flows and 350 thousand UDP flows constituting 7 gigabytes of transferred network traffic. The remaining flows consist of 40 thousand ICMP flows and no IGMP flows. During this time period there have been on the order of ten compromises.

C. Observing Malicious Traffic

In this paper, we visually compare the network flows of a honeynet against the network flows in the NETI@home data. In particular, we make observations to try and answer these three questions:

- What are some of the characteristics of the malicious traffic observed on the Internet?
- How much malicious traffic is seen by end users on the Internet?
- Are there identifiable sources of malicious traffic on the Internet?

The remainder of the paper is organized as follows. First, we will describe some background information on our methods for analyzing the data. Next, we present our findings and compare and contrast the results from the honeynet dataset and the NETI@home dataset. Finally, we discuss some related work and present our conclusions and areas of future work.

II. Network Flow Analysis

In order to compare the NETI@home dataset with the honeynet dataset, we ran a customized version of the NETI@home client on our honeynet data. This yielded flow based statistics of the honeynet data that is in the same format as the NETI@home statistics and is suitable for comparison. In this section, we describe some of the statistics that are provided by the NETI@home client.

The NETI@home client collects statistics for four common transport layer protocols: TCP, UDP, ICMP, and IGMP. Much of our analysis focuses on TCP flows since they make up the majority of the traffic seen in our datasets. However, some data from UDP, ICMP, and IGMP are also presented in our results.

The analysis technique is centered around the concept of a bidirectional flow, based on the commonly used 5-tuple, which consists of the source and destination IP addresses, source and destination ports, and the transport layer protocol. Statistics gathered for each TCP flow include various time measurements, the number of packets sent and received, the source and destination parameters, failure flags, window size measurements, and various other information. Similar statistics are gathered for the flows that are of the other types of transport layer protocols. A full discussion of the statistics gathered can be found in [3].

Each flow has a *local* IP and port number and a *remote* IP and port number. *Local* refers to the host on which the client is running and collecting statistics from. *Remote* refers to the other host in the flow. Therefore, if a NETI@home user with IP x makes a web request to a given IP y, then x would be the local IP and y would be the remote IP. To further clarify, if the same NETI@home user was scanned by IP z, then x would still be the local IP and z would be the remote IP.

There are several sources of bias in our datasets that may skew our results and are worth mentioning. First, an insignificant number of NETI@home users had their clocks misconfigured so we did not include them in the results. Clock synchronization in general is a source of bias. Second, we did not include all IP results from NETI@home users when their privacy was set to high because their IP addresses are unknown. Third, the honeynet dataset is known to be complete; however, the NETI@home dataset relies on the end users to run the NETI@home client to monitor their systems and so may have some incomplete results. Fourth, the NETI@home users must volunteer to run the client, so the data are not a truly random sample of Internet end users. Finally, the honeypots are all on the same network, whereas NETI@home users are spread throughout the Internet.

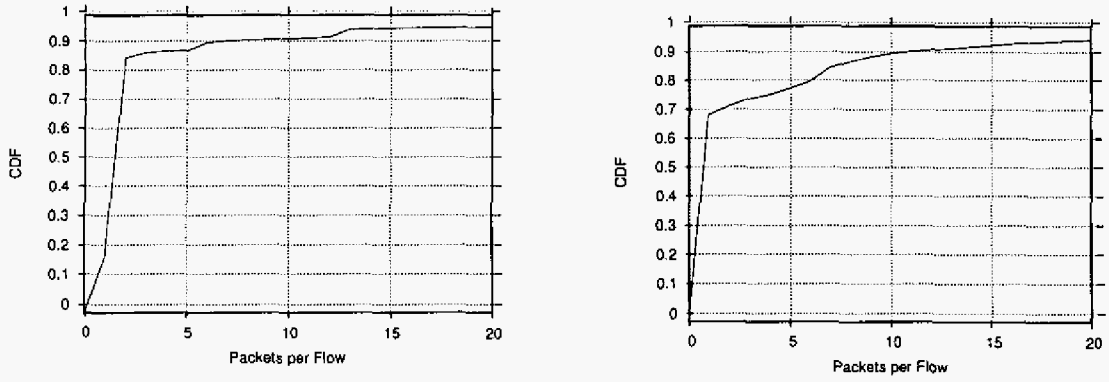
After collecting the flow statistics for both datasets, we created a framework to analyze the data. This framework allowed us to plot various graphs for both datasets for comparison. Below, we present these graphs and discuss our observations.

III. Data Observations

In order to aid in understanding what makes up the majority of the malicious traffic on the Internet we have plotted various metrics for both the honeynet dataset and the NETI@home dataset. The NETI@home dataset represents a mixture of both legitimate/good traffic as well as malicious traffic. The honeynet dataset represents almost entirely malicious traffic. Comparing and contrasting these results can initiate a better understanding of the malicious traffic seen on the Internet.

A. Number of Packets Per Flow

In our first figure we graphed the cumulative distribution function (CDF) of the number of packets for all TCP flows for each dataset. The results are shown in Figure 1. First observe the honeynet curve. One can see two distinct inflection points for packet counts equal to one and two. TCP flows which consist of just one packet most likely con-



(a) Honeynet

(b) NETI@home

Fig. 1. Cumulative distribution function of the number of packets per TCP flow

tain one SYN packet. It is possible to have a single packet flow that is not a SYN packet. For instance, a RST or SYN/ACK packet could be received from a host that received a spoofed connection attempt. We did not observe many flows of this nature.

TCP flows which consist of two packets most likely consist of one SYN and one RST packet or one SYN and one SYN/ACK packet with no final ACK to complete the three-way handshake. Again, there are other combinations of TCP flows consisting of just two packets, but we have not observed many of these combinations. Any TCP flow consisting of two or less packets is a failed connection. On a honeynet, we consider these failed connections to be malicious probes. Therefore, on our honeynet dataset about 87% of all TCP flows can be considered to be probes.

We can contrast the NETI@home CDF with the honeynet CDF and see that about 73% of all TCP flows can be considered failed connections. In the NETI@home dataset, not all of these failed connections are necessarily malicious probe packets as they may be legitimately failed connections. However, it is interesting to note that in terms of number of packets per flow the majority of observed TCP flows for end users are either probes or failed connections.

B. TCP Port Histogram

To better understand what ports and services malicious flows are targeting, we have generated a TCP Port Histogram over time for both the honeynet dataset as seen in Figure 2 and the NETI@home dataset as seen in Figure 3. Each row of points represents one day. The width of the rows span the local TCP ports from 0 to 1024, which are the well known ports [4]. The following formula was used to create the graphs, where i is the intensity value for a given point in a given row:

$$i = \begin{cases} 0 & \text{if } c = 0 \\ 0.75 \cdot \left(\frac{c}{c_{\max}}\right)^{0.45} + 0.25 & \text{otherwise} \end{cases} \quad (1)$$

The maximum number of packets destined to a certain port on one day (i.e. one row in the figure) is denoted $c_{\max}$. A port with a packet count c is then visualized with intensity i according to above formula. If c is zero, the intensity is also set to zero (black). Otherwise, the intensity is chosen to be a value between 25% gray ($i = 0.25$) to white ($i = 1.0$, for the port where $c = c_{\max}$). The exponent is used to boost dark pixels to make them more visible. We choose to represent no activity with dark regions because it provides better contrast for the faint areas of activity.

There are a number of observations to be made from these graphs. Two important characteristics of the figures to observe are the horizontal lines and the vertical lines. First, the horizontal lines represent port scans. Port scans are often malicious in nature as an attacker will generally use a port scan against a target in order to determine that target's weaknesses. In the honeynet data, a number of port scans can be seen over time, but the NETI@home dataset shows a significantly denser number of port scans seen over time. This appears to be intuitive as there are an order of magnitude more NETI@home users, which are distributed across the Internet both topologically and geographically, than there are honeypots in our dataset. Some factors that would decrease the number of port scans seen by NETI@home end users include firewalls, NATs, or other similar configurations. Even with these factors, some NETI@home users are seeing similar port scans as seen on our honeynet.

Another interesting observation is that there are a number of different types of scans seen. At least four different port scans are easily distinguished visually in the honeynet data as denoted by the letters $A - D$, and similar scans are

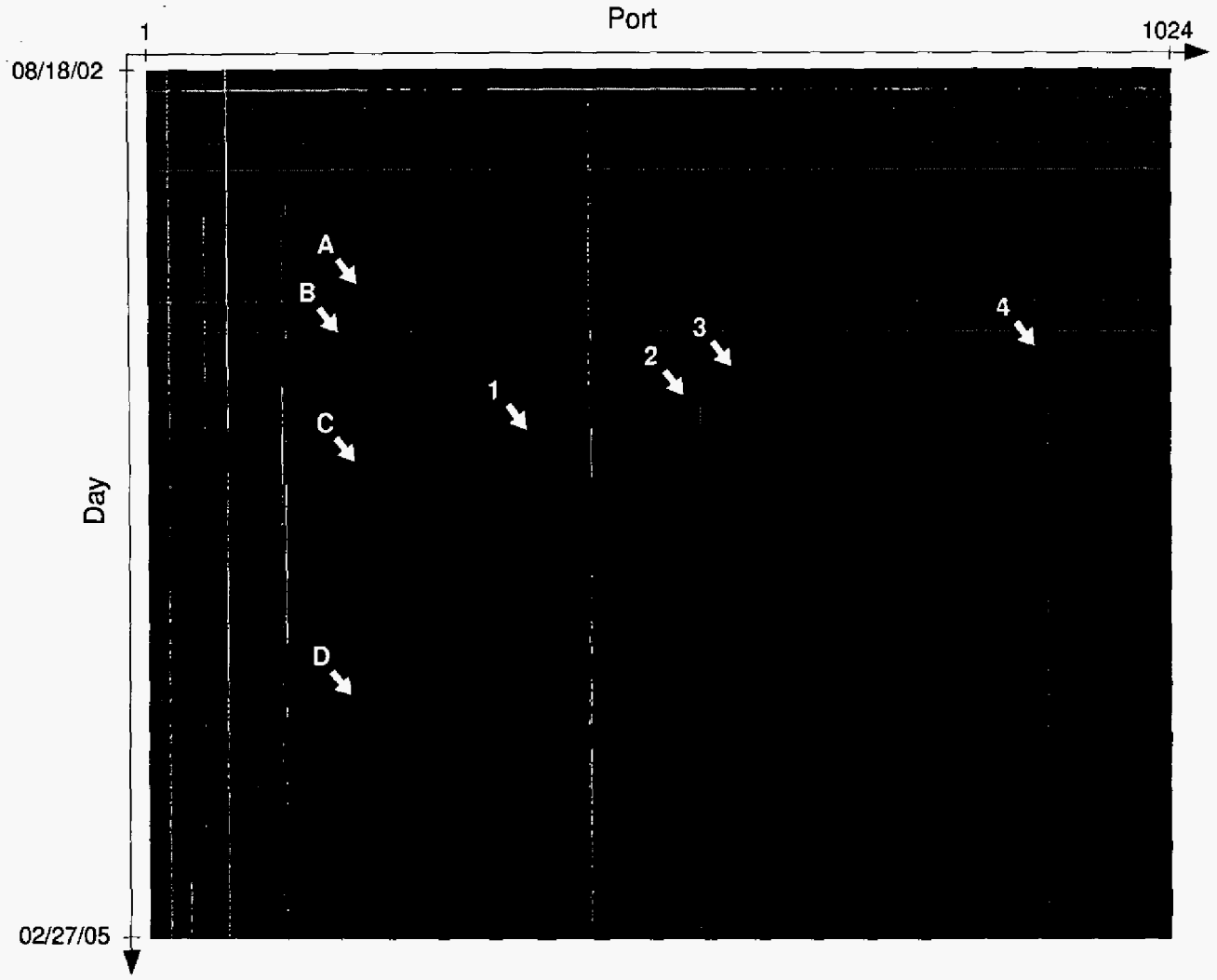


Fig. 2. Honeynet TCP Port Histogram

observed in the NETI@home data. The most naive port scan will scan all ports (B). The more sophisticated port scans will skip ports that are of little interest (A, C, and D). There are a number of widely available port scanning tools, which offer various options for the scanning algorithm [5,6].

One interesting difference seen in the horizontal lines in the NETI@home dataset are the stair step lines from approximately port 512 through 1024. Since the user that reported these flows was within the Georgia Tech network and used a low privacy level, we were able to determine what caused the stair step lines. An administrative machine within the Georgia Tech network was scanning ports 512 through 1024 over the course of several days. The algorithm consists of dividing the ports into a number of ranges and scanning one range each day. The source of the scanning was a machine used to help secure the network and so was altruistic. Therefore, we do not consider these scans to be malicious in nature.

The second interesting aspect to observe in these graphs are the vertical lines. The vertical lines represent ports that have continual traffic over large time scales. Looking at the honeynet graph from left to right, the most prominent TCP ports with continual traffic are 22 (ssh), 80 (www), 135 (Microsoft Windows Service), 139 (Microsoft Windows Service), and 445 (Microsoft Windows Service). Most of these ports have been a target of one or more worms in the past in addition to legitimate traffic.

There are a number of other vertical lines that are not as prominent in the honeynet dataset as seen in Figure 2. The

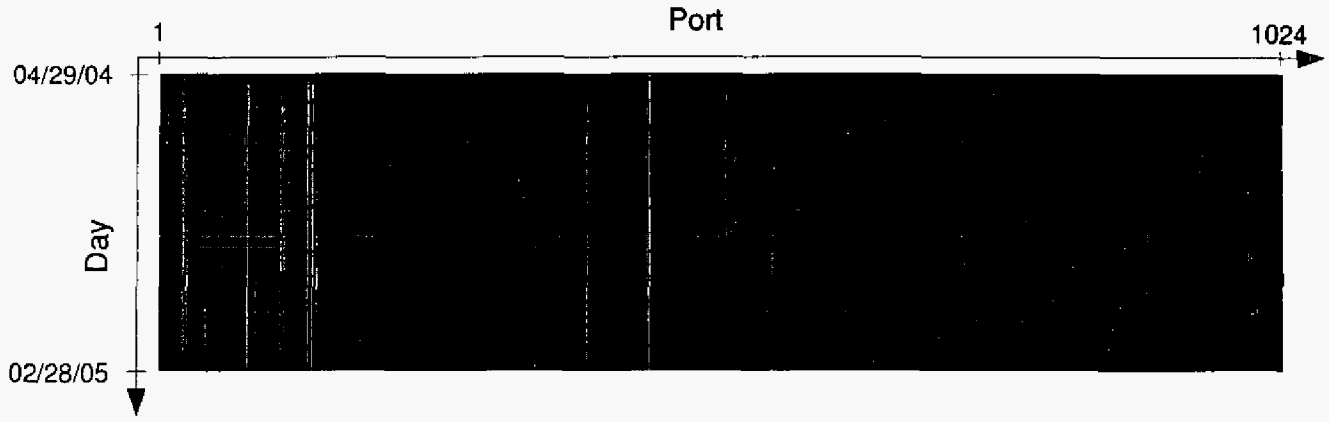


Fig. 3. NETI@home TCP Port Histogram

vertical line denoted by '1' is LDAP traffic and was only seen for a short period of time. The line denoted by '2' represents traffic seen from the real time service protocol worm. The traffic at '2' is particularly interesting in the honeynet dataset. One can notice a bright burst of traffic starting on the worm release date that continues with intensity over the course of the next several days. After a number of days, the worm traffic slowly fades out as the infected machines are repaired. However, trailing effects of the worm can be seen from the point of release until the end of the dataset, which is over the course of more than a year. Therefore, we see lingering worm traffic exists on the Internet for long periods of time after the initial release date.

The line denoted by '3' represents traffic seen from the blaster worm as seen in Figure 2. This line also continues on for a long period of time, although its characteristics are not as distinguishable as the real time service protocol worm. In the honeynet data, it is not clear why traffic is seen at the line denoted by '4' at port 901. This may be traffic targeting an old Trojan port, RealSecure's management port, or Samba/SWAT on RedHat Linux based boxes. It is interesting to note that these trends seen in the honeynet data are repeated in the NETI@home data in addition to the legitimate traffic as seen in Figure 3. Although, it is difficult to distinguish between legitimate traffic and worm traffic in the NETI@home dataset.

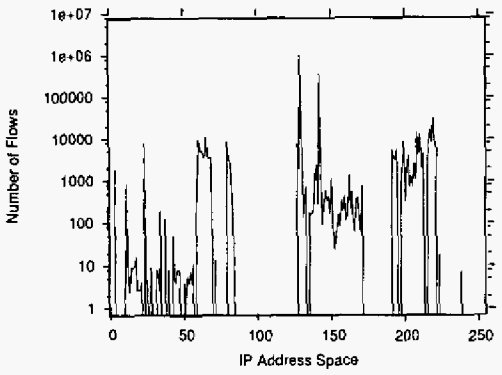
C. IP Address Space

The graphs in Figure 4 show where the traffic is coming from or going to within the entire IP address space. The IP address is divided into 256 buckets based on the first byte of the IP address. Figure 4(a) shows the honeynet graph. It is clear that certain portions of the address space have seen zero activity on the honeynet. These portions correspond with unallocated addresses as listed in the whois database. Given that there are no flows from most of these spaces to the honeynet, we conclude that there are not many spoofed IP packets coming from unallocated IPs to our honeynet. Further, either the number of packets with spoofed IP addresses coming to our honeynet is low or they are intelligently designed.

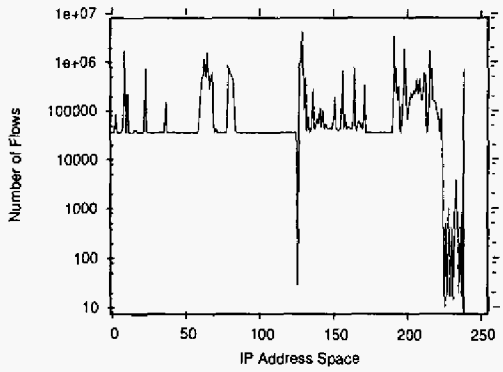
The NETI@home dataset has an additional baseline of traffic seen across most of the address range as seen in Figure 4(b). Further investigation found that this baseline is caused by one or more NETI@home users sending out a large number of TCP flows to TCP port 445 over a short period of time. We are unsure how many users were reporting these results due to privacy settings. Figure 4(d) shows the number of flows to TCP port 445 versus the IP address space. There is clearly a horizontal line across the majority of the IP address space, which suggests that the NETI@home user or users were randomly scanning the IP address space on TCP port 445. The nature of this scanning may have be malicious in nature. For example, the user may have been infected with a worm as there have been worms that target TCP port 445. However, we cannot conclude for certain that the traffic was malicious in nature.

In Figure 4(d), there is a small increase in traffic at bucket number 10. This is probably due to local 445 traffic on private 10.0.0.0/8 networks. Similarly, there is an increase in traffic at bucket number 192. This increase would be due to local 445 traffic on private 192.168.0.0/16 networks. The sharp drop in traffic at bucket number 127 is due to the fact that the 127.0.0.0/8 network is the dedicated localhost network. Finally, the upper ranges of the IP address space did not see any scans. These ranges contain multicast, experimental, and other types of allocations.

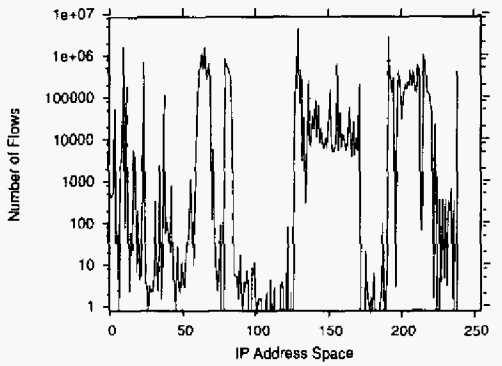
To better compare the NETI@home data with the honeynet data, we graphed the NETI@home dataset filtering out traffic to TCP port 445 as seen in Figure 4(c). Com-



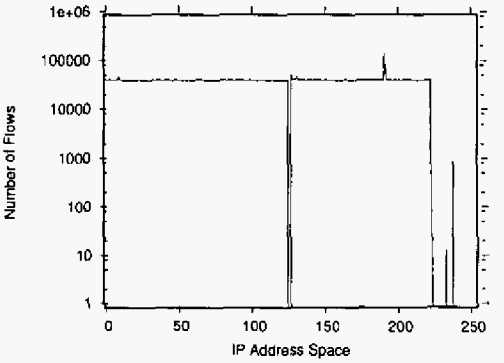
(a) Honeynet (all IP flows)



(b) NETI@home (all IP flows)



(c) NETI@home (no TCP port 445 flows)



(d) NETI@home (only TCP port 445 flows)

Fig. 4. IP address space distribution by number of flows

paring Figures 4(a) and 4(c), one can notice a striking similarity between the NETI@home data and the honeynet data. Some differences in the NETI@home data include traffic to the multicast range and some traffic in the unallocated ranges. However, visually the two graphs have notably similar shapes.

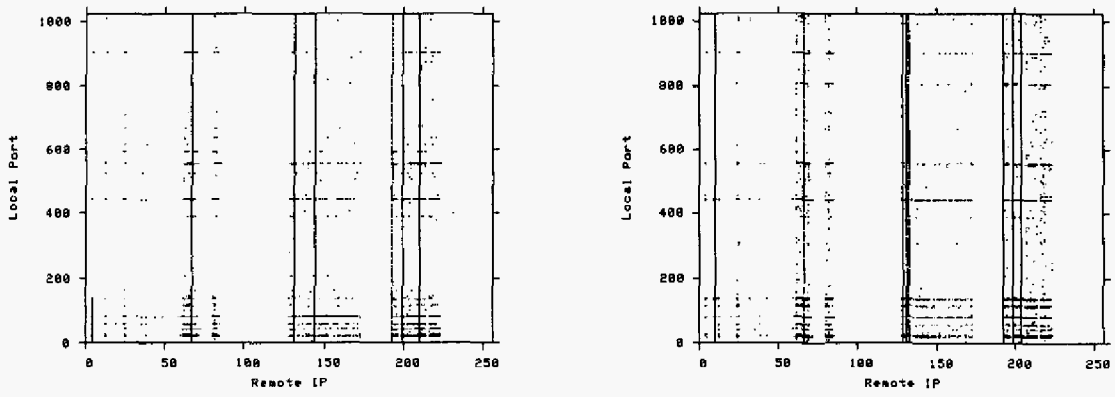
Based on our observations of the IP traffic seen relative to IP address space, we note a possible algorithm for detecting suspicious machines. In previous work, we showed that a honeynet can be used to find compromised machines on large enterprise networks by marking any machine on the enterprise that attempts to connect to the honeynet as suspicious [7]. An extension that we draw from these graphs is that any machine attempting to connect to an unallocated IP address should be considered suspicious and may be compromised.

A graph of the remote IP versus local port for both datasets can be seen in Figure 5. Again, we only plot the well known TCP ports. In these graphs, one can see that remote IPs that appear in the flows are spread across the allocated IP spectrum, and again there is little traffic in the unallocated ranges, even in the NETI@home data. Based on these graphs, we observe that scans come from across the entire allocated IP address space.

IV. Related Work

Much work has been accomplished on measuring Internet statistics. The Cooperative Association for Internet Data Analysis (CAIDA) was founded in order to provide "tools and analyses promoting the engineering and maintenance of a robust, scalable global Internet infrastructure." [8]. CAIDA examines all aspects of the Internet including topology, routing, performance, and security. Much of the CAIDA measurements and results focus on macro–Internet observations while we present micro–Internet observations as seen by end hosts.

There has also been much research on Internet worms. Work has been accomplished on characterizing and looking at the trends of various worms [9, 10]. Further, a detailed study of the spread time, algorithms, and damage caused by recent worms has been conducted. For example, Shannon et. al. give an in depth look at the Witty worm in [11], and Moore et. al. give an in depth look at the Slammer worm in [12]. We see both of these worms in our dataset



(a) Honeynet (b) NETI@home

Fig. 5. Remote IP address and contacted local TCP port

and data shows that their lingering effects are still active.

Various schemes for measuring Internet activity have been designed and implemented. CAIDA uses a network telescope, which consists of a full /8 network in order to observe worms, DoS attacks, network scanning, and other malicious activity [13]. SANS recently started the Internet Storm Center (ISC) in order to provide users and organizations with warnings against possible new threats seen on the Internet [14]. The NETI@home dataset focuses on end user statistics, while the honeynet dataset can be considered similar to the network telescope data, with the exception that live hosts will respond to probes.

V. Conclusions and Future Work

We used a number of methods to analyze network flows over time for NETI@home data and honeynet data. In both datasets, the majority of the TCP flows were failed connections. In the honeynet dataset, these flows were malicious in nature. The NETI@home dataset has a smaller percentage of TCP flows that were failed connections, and these flows were not necessarily malicious in nature.

The majority of the traffic seen in the honeynet dataset consists of port scans and worms. We observed that the outbreak of a new worm will linger on for more than a year after the release date. Similar patterns were observed in the NETI@home data, although it is difficult to distinguish between malicious and legitimate traffic.

We also found that port scanning was seen by NETI@home users and honeynet machines regularly. By using our technique of a TCP port histogram, we were able to observe an altruistic port scan of NETI@home users that slowly scanned the ports over the course of several days. Some of the malicious port scanning patterns observed in the honeynet dataset were also observed in the NETI@home dataset.

We found that both datasets showed similar flow distributions across the IP address space. In the NETI@home dataset, however, a small number of users were scanning most of IP address space in a random fashion on a TCP port that is the target of recent worms. Finally, for both datasets, the source of malicious and legitimate traffic comes from across the entire allocated IP address space. We did not observe significant malicious traffic or legitimate traffic coming from the unallocated IP address space.

There are a number of future directions to research. We intend to do a formal statistical correlation between the honeynet data and the NETI@home data to draw more definitive conclusions. There are numerous other network statistics that can be compared such as TTL values, window sizes, checksum errors, and so forth. The analysis of these areas of research will be conducted in future work.

References

[1] "NETI@home." `http://www.neti.gatech.edu`, March 2005.
[2] "Georgia Tech honeynet research project." `http://www.ece.gatech.edu/research/labs/nsa/honeynet.shtml`, March 2005.
[3] C. R. Simpson and G. F. Riley, "NETI@home: A distributed approach to collecting end-to-end network performance measurements," in *PAM2004 - A workshop on Passive and Active Measurements*, April 2004.
[4] J. Reynolds and J. Postel, "Assigned numbers," October 1994. *RFC 1700*.
[5] "nmap." `http://www.insecure.org/nmap/`, March 2005.
[6] "nessus." `http://www.nessus.org/`, March 2005.
[7] J. Levine, R. LaBella, H. Owen, D. Contis, and B. Culver, "The use of honeynets to detect exploited systems across large enterprise networks," in *Proceedings of 4th IEEE Information Assurance Workshop*, (West Point, NY), June 2003.
[8] "Internet measurement infrastructure - caida." `http://www.caida.org`, March 2005.
[9] D. M. Kienzle and M. C. Elder, "Recent worms: a survey and trends," in *WORM'03: Proceedings of the 2003 ACM workshop on Rapid Malcode*, pp. 1–10, ACM Press, 2003.
[10] N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, "A taxonomy of computer worms," in *WORM'03: Proceedings of the 2003 ACM workshop on Rapid Malcode*, pp. 11–18, ACM Press, 2003.

[11] C. Shannon and D. Moore, "The spread of the witty worm," *Security & Privacy Magazine*, vol. 2, no. 4, pp. 46–50, 2004.
[12] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the slammer worm," *Security & Privacy Magazine*, vol. 1, no. 4, pp. 33–39, 2003.
[13] "Telescope analysis - caida." `http://www.caida.org/analysis/security/telescope`, March 2005.
[14] "Sans - internet storm center." `http://isc.sans.org/`, March 2005.