Remote Fingerprinting and Multisensor Data Fusion

Samuel Oswald Hunter
Dept. Computer Science
Rhodes University
Grahamstown, South Africa
Email: sam@rootentropy.co.za

Etienne Stalmans
Dept. Computer Science
Rhodes University
Grahamstown, South Africa
Email: g07s0924@campus.ru.ac.za

Barry Irwin
Dept. Computer Science
Rhodes University
Grahamstown, South Africa
Email: b.irwin@ru.ac.za

John Richter
Dept. Computer Science
Rhodes University
Grahamstown, South Africa
Email: j.richter@ru.ac.za

***Abstract*—Network fingerprinting is the technique by which a device or service is enumerated in order to determine the hardware, software or application characteristics of a targeted attribute. Although fingerprinting can be achieved by a variety of means, the most common technique is the extraction of characteristics from an entity and the correlation thereof against known signatures for verification. In this paper we identify multiple host-defining metrics and propose a process of unique host tracking through the use of two novel fingerprinting techniques. We then illustrate the application of host fingerprinting and tracking for increasing situational awareness of potentially malicious hosts. In order to achieve this we provide an outline of an adapted multisensor data fusion model with the goal of increasing situational awareness through observation of unsolicited network traffic.**

***Index Terms*—remote fingerprinting, data fusion, situational awareness**

I. Introduction

REMOTE host fingerprinting involves the inferred identification of software, services and hardware of devices based on observed characteristics that match known signatures. The characteristics of the device could be observed passively over a network, however in most cases probes are sent to the device in order to elicit a response. While host fingerprinting is most commonly used to map attack vectors during the reconnaissance phase of an attack, we propose the application of fingerprinting techniques towards hosts on the Internet that have malicious intent. This allows one to not only learn more about the characteristics of compromised hosts, but also to track these nefarious hosts in dynamic IP address space. When a previously observed host is re-identified it allows us to construct a history of past discrepancies which provides insight into the characteristics and behavior of these hosts.

With the exception of traffic anomalies such as those caused by miss configured hardware or software services [1], unsolicited network traffic may be considered as either potentially malicious (PM), truly malicious (TM) or as a result of malicious (RoM) activity. We classify PM activity as traffic consisting of probes from host discovery scans, port scans or vulnerability scanning; these types of scans are often produced by tools such as Nmap[1] and Nessus[2]. TM activity consists of exploitation attempts against a system, such as an entity attempting to exploit the MS08-067 vulnerability [18] on a host. Truly malicious traffic is often caused by worms exploiting a system or a malicious hacker using a tool such as Metasploit[3]. RoM activity, also known as backscatter, consists of traffic received as the result of certain types of Distributed Denial of Service (DDoS) attacks [12]. When compromised hosts (bots) are used in a DDoS attack they create spoofed source IP addresses for the duration of their attack. In certain cases the spoofed IP address will overlap with the address range of our monitoring sensors enabling the detection of backscatter or RoM traffic from the attack.

The vast majority of unsolicited traffic can thus be seen as originating from nefarious hosts who have some form of malicious intent. By monitoring this traffic we are able to extract existing attack methods and detect new techniques [1][15]. This type of traffic is, for the most part, obfuscated by legitimate traffic on production systems. There are, however, two techniques that are used to detect and capture unsolicited traffic: network telescopes and honeypots. This paper introduces a normative method of discovering hosts that have potential for producing PM and TM traffic as well as instigating RoM traffic through the enumeration of bots associated with a Fast-flux domain in section IV-B.

This work proposes four categories to better define remote host fingerprinting; software, physical, associative and behavioral. Data observed by unsolicited traffic monitoring sensors undergo a data fusion process which is defined by our adapted model of the multisensor data fusion model initially described by Waltz [20]. This model was later adapted for use with cyber situational awareness [4] and distributed intrusion detection systems [3]. By creating our own adaption of a multisensor data fusion model we show how it can be used to gain

[1] Nmap - Port Scanner, http://nmap.org/

[2] Nessus - Vulnerability Scanner, http://www.tenable.com/products/nessus

[3] Metasploit - Penetration Testing Software, http://www.metasploit.com/

situational awareness with regards to potentially malicious hosts on the Internet.

Two new techniques, latency multilateration and associative fingerprinting, are introduced in this work. It was found that the latency-based technique, as a form of physical fingerprinting, showed promising results as a supporting metric; however due to some sporadic results it cannot be used reliably on its own. Associative fingerprinting assumes a membership based relationship between observed hosts and some physical or logical entity. A 1000 compromised hosts from the Kelihos/Hlux botnet [14] were enumerated: characteristics of these hosts were analysed and this research shows how they provided insight into unique attributes that can be used to associate hosts with a given botnet.

The topic of host-tracking, excluding application level tracking, has seen very little research. Host tracking is a best-effort attempt relying on probabilities based on repeatedly observed characteristics. It is, however, of great interest to monitor behavior of hosts in dynamic IP address space over time. Our research based on host tracking, may also have far-reaching implications in fields such as computer forensics and information warfare. In this paper we provide several host-defining metrics, which, while used as separate fingerprints might not provide a unique representation of a host but will, when used as a collective, fingerprint multiple, inherently different attributes of a device. This then increases the probability that the device can be uniquely identifiable.

By making use of multiple host-defining metrics, obtained through remote fingerprinting and the adapted data fusion model for monitoring unsolicited network traffic, we create situational awareness that can provide deeper insight into the representation of malicious hosts on the Internet. The observation of their behavior and the ability to keep track each of these hosts over time may allow us to identify the most prolific sources of malicious behaviour on the Internet and possible infer their intentions against our own networks.

Section II will discuss our adapted multisensor data fusion model and the various data sensors of which it comprises. In Section IIIwe define the four categories of remote host fingerprinting which can provide metrics for host tracking. Section IV provides a discussion and the results of the two fingerprinting techniques; latency-based multilateration and associative fingerprinting.

II. Multisensor Data Fusion

Multisensor fusion techniques combine data from multiple sensors and related information from data stores to achieve greater confidence and accuracies in results and to produce inferences that would not have been as reliable had they been produced from from a single sensor's observations [6][10][20]. Data fusion involves the hierarchical transformation of data from multiple and often heterogeneous sources. The transformation process is coupled with a decision making and inference based classification of the data characteristics. The context of the observed environment and the relationships between entities therein are also of importance during the fusion process.

Traditional military command and control (C2) systems would make use of multiple field sensors to observe and measure data primitives such as radiation, acoustic and thermal energy, nuclear particles and other observable signals. The sensors would then be used in a data fusion process in order to correlate, aggregate and associate the data in order to assist in an automated decision making process or support system. While in principle the process of multisensor data fusion would provide significant advantages over the observations of a single data source, in [7], Hall states that besides the statistical advantage gained by combining same-source data, the use of multiple types of sensors might - in practice - produce worse results than could be obtained by making use of a single, more appropriate sensor for the application. Hall further explains that these sub-optimal results are caused by an attempt to combine accurate data with inaccurate or biased data.

We propose an adaptation of the generic data fusion model by Waltz [20], while noting the work done by Bass [3][4] in which the generic model was adapted and a framework for next generation distributed intrusion detection systems was outlined. The model is adapted by including active response mechanisms to events, handling unsolicited network traffic as input and generating situational awareness as output. Active reconnaissance of objects modeled in the data fusion process allows for a more detailed and effective representation of those objects. The active reconnaissance forms part of remote host fingerprinting. The remainder of this section discusses the adapted fusion model in greater detail and introduces a variety of data sensors for the capture of unsolicited traffic.

A. Data Sensors

Network telescopes[12], also known as darknets[5] are, in their most simplest form a service operating over an unused but publicly accessible IP range. The service's sole purpose is to log all traffic destined towards it. The use of network telescopes in the assessment of potentially malicious traffic as a result of unsolicited communication has gained popularity over the last decade for its ease of deployment and low cost investment. Typically a honeypot would emulate one or more vulnerable services in an attempt to lure malicious entities into interacting with it.

Traffic observed by network telescopes and honeypot sensors would be disseminated and entered into the data fusion process through the use of a messaging framework [9]. The messaging framework was implemented using the Advanced Message Queuing Protocol (AMQP) with a RabbitMQ[4] broker. A publish/subscribe model was chosen as it would allow for scalable data exposure and simplified distribution of work amongst fingerprinting modules. Modular data exposure is achieved through the use of JSON encoding. Two of the main advantages of using a message queuing service with a publish/subscribe model is the near realtime data exposure and the scalability inherent with a publish/subscribe data distribution model.

[4] RabbitMQ - Messaging Framework, http://www.rabbitmq.com/

The Fast-Flux Botnet Enumeration and Online Scraping tools that were developed for this research are responsible for producing support-based information, as apposed to the primitive data observations of the network telescopes and honeypots. This information assists in decision making and threat analysis at later stages of the multisensor data fusion process.

1) Network Telescopes: Previous research concerned with the analysis of network telescope traffic has been successful in observation, identification and tracking of Distributed Denial of Service attacks [12], worm propagation [8][16] and network scanning [2]. The function of network telescopes - to capture traffic destined for unused address space - is particularly attractive as no legitimate traffic should exist on an unused range. By capturing only unsolicited traffic, network telescopes show that they are well suited to data capture, assisting in understanding the state of illegitimate and potentially malicious traffic on the Internet. As a result,network telescopes where chosen as one of the data sensors to be incorporated with the multisensor data fusion. Network telescopes are used to provide primitive observations. Metrics from the traffic that are important are listed in Table I.

Timestamp	Sequence Number
Source IP	Destination IP
Source Port	Destination Port
Protocol	Flags
UDP Payload (in some cases)	

Table I
ATTRIBUTES OF INTEREST FROM NETWORK TELESCOPE SENSOR TRAFFIC.

2) Honeypots: A honeypot is a device or service that operates in a network, designed to detect various forms of malicious interaction that initiate connections with it. In doing so honeypots can serve as a defensive mechanism by acting as decoys for an attacker, but also as a valuable resource when observing nefarious traffic. A Dionaea[5] honeypot was used as one of the sensors. Data observed by the Dionaea honeypot was exposed through the XMPP protocol for realtime dissemination. The honeypot provided primitive observation. Metrics from the honeypot data are listed in Table II.

Attributes	
Timestamp	Attack Description
Source IP	Destination IP

Table II
ATTRIBUTES OF INTEREST FROM HONEYPOT SENSOR DATA.

3) Fast Flux Botnet Enumeration: Fast-Flux is a technique used to rapidly update DNS information. While it can be used for legitimate purposes such as load sharing, it has become popular with resilient botnets [13]. It has been shown that "bot herders" make use of Fast-Flux DNS techniques to host content within a botnet. This allows address mapping to constantly shift between different bots, making it very difficult to shut down [13]. Bots within the botnet relay content back to a central server, also known as the "mothership" [13]. Our own research into the Kelihos/Hlux botnet confirms this through the detection of an Nginx[6] web server on each of the bots found. This web server can act as a reverse proxy to forward content from a central server. In order to analyse the characteristics of a compromised host (bot) that belongs to a botnet, a tool to harvest IP addresses of hosts belonging to a Fast-Flux botnet was developed. The data sample was obtained by harvesting 1000 unique IP addresses from the Kelihos/Hlux botnet during March of 2012. Figure 1 shows that the majority of IP addresses were returned between 5 and 22 times, while the most returned IP address was recorded on 77 different occurrences.

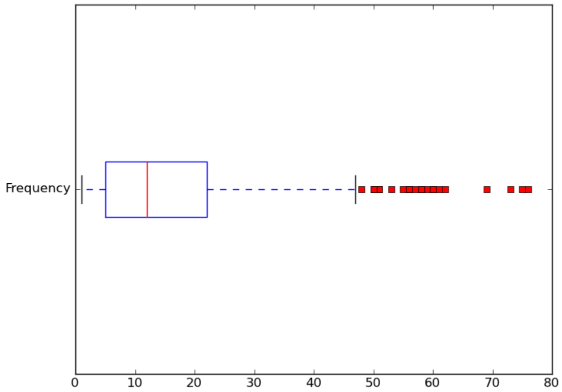

Figure 1. Frequency of IP addresses that where returned during the Kelihos/Hlux botnet enumeration. The x-axis shows the number of times a given IP was returned during the enumeration process for discovering a 1000 unique IP addresses.

As mentioned earlier, the Fast-Flux Botnet Enumeration tool developed for this research does not act as a primitive data sensor: instead it provides supporting information to be used in association with entities observed by the network telescope and honeypot sensors with botnet or bot-like characteristics. The tool takes a Fast-Flux domain as input and a target number of hosts to enumerate. The Fast-Flux Botnet Enumeration tool will then find, through multiple DNS requests, the requested number of hosts, stopping once a certain number of attempts at finding a new host has been reached. The IP addresses of the hosts are then enumerated by performing a basic port scan on each of the hosts and a HTTP GET request to port 80 is performed. Results are shown in Section IV-B.

4) Online Sources: While strictly speaking the Online Scraping Tools and Fast-Flux Botnet Enumeration tool do not constitute data sensors in the adapted multi-sensor data fusion model, they are listed in this section for brevity. They form part of Level 4 (Resource Management).

There are numerous websites on the Internet that maintain databases of information concerning malicious activity on the Internet,as well as data regarding the sources of malicious activity. By making use of these sources it is possible to produce rich datasets containing information such as past

[5] Dionaea - Honeypot, http://dionaea.carnivore.it/

[6] Nginx - HTTP and reverse proxy server, http://nginx.org/en/

discrepancies of hosts, the time at which malicious activity was first observed and the kind of malicious traffic the host produced. These sources of information, while useful when used as supporting data, can also be used as a data source for finding malicious hosts to analyse. These sites contain large quantities of information on historical malicious activity and are updated on a regular basis, ranging from continuously to daily. We made use of projecthoneypot.org in order to identify the nature of malicious activity from IP addresses that cross-referenced with their database.

B. Fusion Model

Inputs to the data fusion process include the sensor data outlined in Section II-A, *priori* data, supporting data collected in real-time through reconnaissance techniques such as port scans, and Operating System identification. Remote fingerprinting support data is where the adaption of the generic multi-sensor data fusion model becomes most apparent. The process of multisensor fusion is concerned with the adaption, correlation and analysis of data from different sources in order to evaluate a situation or event and then assist in the decision-making process or initiating direct action. Figure 2 illustrates the adapted model for monitoring unsolicited network traffic to produce situational awareness.

Unsolicited traffic sensors provide raw data, in the form of network packets from network telescopes and processed data in the form of events registered by honeypots. This data forms the unsolicited traffic input into the data fusion model. The multi-sensor data fusion model provides an abstract representation of data aggregation, alignment and analysis through the use different levels of processing. These levels form a hierarchy of processing that data undergoes in order to create knowledge. Level 0 data refinement was originally concerned with the calibration of equipment, sensor adjustment and filtering of data [4] such as out-lier removal.

- **Level 0 (Data Refinement)** consists of determining the validity of the source IP address by determining if the address falls in bogon address space.
- **Level 1 (Object Refinement)** is the processing of received data ; aligning the data to a common frame of reference such as timestamp, sequence numbers and location of IP address through ASN.
  - This includes correlation of data, such as packets with the same source IP address. Data is also classified at this level as potentially malicious traffic, truly malicious traffic or response of malicious traffic.
  - This Level is concerned with the creation of an object entity and the association of characteristics to that entity in order to create an Object Base.
- **Level 2 (Situational Refinement)** provides situational knowledge regarding the object base after it has been aligned, correlated and analysed.
  - At this level, aggregated sets of objects may be detected by examining their co-ordinated behavior. An example of this could include the observation of backscatter traffic from a DDoS attack or the traffic from an actual DDoS attack on some portion of the sensors. The Situational Refinement process takes into consideration the various sensors involved, as well as the information provided by Level 4 (Resource Management), which constitutes any additional real-time and *priori* data retrieval such as the use of fingerprinting techniques and searching for past discrepancies associated with characteristics of the Object Base. Characteristics of objects that are inspected at at this level include: common points of origin (source ip or address range), protocols, common targets and attack rates.

It is important to note that unlike the traditional application of multisensor data fusion, data will often only be observed by a single sensor or, alternatively, be observed by multiple sensors but at disjoint points in time. As an illustration, consider a host infected with an SSH worm. The worm scans network ranges in order to find further hosts to infect, and might scan a honeypot first and a day later reach one of the network telescope sensor ranges. The Level 4 (Resource Management) process resolves this disjoint information. This part of the model is used to collect additional information and refer back to older data in order to correlate characteristics. Continuing with this example:

- **Level 3 (Threat Assessment)** could make use of the time difference between detection of the SSH worm scanning the honeypot and the network telescope to infer the rate at which this host is scanning network ranges. The Threat Assessment process is concerned with the possible threats that unsolicited traffic might pose, as well as the implications of these threats. Data from this level is combined with information from Level 2 to construct a Situation Base.
- **The Situation Base** represents aggregated information from a single entity or multiple entities, combining information from all the levels in the multisensor data fusion.

This represents an event that constitutes a small subset of the growing situational awareness that ultimately corresponds to the available intelligence on a malicious host demographic.

III. Remote Fingerprinting

The process of fingerprinting involves the enumeration of attributes from devices or services and the correlation of these attributes with a list of known signatures of potential devices or services. One of the objectives of this research is to establish, with some certainty, a unique fingerprint of a host so that we would be able to identify the host again at different points in time and maintain an account of its activity. As not all hosts connected to the Internet make use of static IP addresses, this is a non-trivial process: the vast majority of residential and mobile connectivity is achieved through the leasing of dynamic IP addresses from an Internet Service Provider. While it is impossible to conclusively discover the identity of a host without physical access to that device, we attempt to fingerprint and identify a host with as much certainty as possible. The ability to keep track of a host in dynamic IP address space is applicable in the fields of computer forensics and information warfare. Where previously

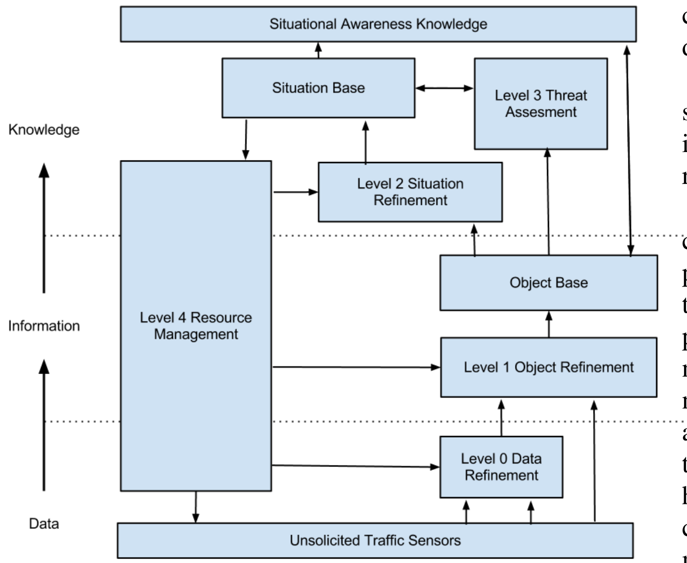
Situational Awareness Knowledge
Situation Base
Level 3 Threat Assesment
Knowledge
Level 2 Situation Refinement
Object Base
Information
Level 4 Resource Management
Level 1 Object Refinement
Level 0 Data Refinement
Data
Unsolicited Traffic Sensors

Figure 2. Unsolicited network traffic fusion for advanced situational awareness adapted from [20][3].

this could only be effectively achieved through the deployment of "call home" code on that host, this is now possible - to a degree - through a holistic fingerprinting process, where multiple heterogeneous remote fingerprinting techniques are combined to create a logical and unique representation of the host. The remainder of this section outlines four unique categories of remote fingerprinting.

It should be noted that it is impossible to remotely fingerprint a host with complete confidence. Fingerprinting relies on traffic from a host, captured either passively or as a result of probing. While it might be unlikely that the host is attempting to interfere or even be aware of the fingerprinting process, research [17] has been performed and applications have been developed for just this purpose. It is assumed that the vast majority of hosts on the Internet are not attempting to subvert our remote fingerprinting attempts.

A. Four Categories of Fingerprinting

In order to uniquely identify a host, the combination of multiple fingerprinting techniques that measure, observe and determine heterogeneous characteristics of a host are required. The four categories of fingerprinting that have been identified are software characteristics, physical characteristics, association/affiliation and behavioral patterns. These four categories represent host defining metrics that, when measured correctly, are capable of creating a unique logical representation of a host that can be used for unique identification.

There exist two distinct types of fingerprinting: active and passive. Passive fingerprinting is undetectable but less reliable. This is achieved by inspecting packets streams either through a Man-in-the-Middle attack, through reflected traffic such as backscatter, or through incoming unsolicited traffic. A popular passive fingerprinting tools is p0f[7]. p0f inspects packets and compares the characteristics of packets to known signatures of different Operating Systems.

The process of active fingerprinting uses probes that are sent to a target in order to elicit a response. This response is then analysed to determine the subset of the characteristics representing the device.

1) Software Characteristics: Software characteristics include metrics such as Operating System and OS version, open ports, services running on those ports and the versions of those services as well as unique configurations. The three most popular tools used for fingerprinting within this category are nmap, p0f and hping[8]. This research made extensive use of nmap and p0f. Enumerating the software characteristics of a host may provide insight into the purpose and intent of that host. It also maps an attack vector when targeting that host with offensive capabilities, as it provides information concerning the possible Operating System and services that may be vulnerable to attack. In certain instances it may also provide a unique value used to identify the host. For example, it may obtain the public SSH key of the host if it is running an SSH service.

2) Physical Characteristics: The physical characteristics of a host are represented by attributes such as hardware and geographic location in the world. A previous study [11] into clock skews for remote fingerprinting of hardware has shown promise. In order to fingerprint a device, the study exploits microscopic deviations in device hardware allowing the authors to uniquely identify devices even when behind a NAT or firewall [11]. Another physical characteristic of a host is the host's MAC address, which represents a unique hardware address of a Network Interface Card(NIC). There are, however, two challenges in obtaining the MAC address of a host: the MAC address cannot be obtained across subnets and it is trivial to spoof a MAC address. The third metric for physical attributes is geographic location. This can be determined by the IP address of a host, however this is subject to potential inaccuracy. For our research we have built further on the concept of geographic locality to fingerprint a host by implementing latency based measurement, discussed in Section IV-A.

3) Association: We define associative fingerprinting as the identification of a relationship, association or affiliation between a host and some physical or logical entity. While this approach is normative, to our knowledge it has not been formally defined or used as a fingerprinting technique. An Internet Service Provider is an example of a host/entity relationship: in it's simplest form an ISP is responsible for providing Internet connectivity and an IP address to a device. This relationship is, however, not absolutely unique and can thus not be used as a fingerprint that defines a unique characteristic of a host. It is, in fact, slightly more unique than a dynamic IP address as it has been leased. Botnets provide another example of a host/entity relationship: the host has a logical association with an entity: the botnet. This can be logically extended by associating the host to a specific botnet. The method by which we identify a host as a bot and then associate that host with a specific

[7] p0f - Purely passive fingerprinting, http://lcamtuf.coredump.cx/p0f3/

[8] Hping - Network Scanner, http://www.hping.org/

botnet is discussed in Section IV-B.

4) Behavioral: We consider behavior attributes of a hosts to include time spent online or offline and average congestion or bandwidth available to that host. The behavioral characteristics of a host are the most challenging to monitor and collect remotely. Without authorization to run monitoring tools on the host, the metric becomes a best effort attempt coupled with inferences. The following methods will be explored and expanded on in future work:

Connection time monitoring requires one of two prerequisites - a static IP address or a resolvable domain name. If the host is reachable through either of those methods it is possible to send periodic probes and construct a behavioral pattern for the time that host spends connected to the Internet from the responses.

Bandwidth availability and load can be determined through probes such as ICMP ping requests - monitoring the response times provides an indication of network activity. If, for example, the responses take longer from 17:00 to 18:00 GMT every day, we could infer that the host is performing backups to an off-site location. This would constitute a valid behavior attribute that, together with other fingerprinting techniques, could be used to uniquely represent a host.

IV. Latency Based and Associative Fingerprinting

The ability to track (or "re-identify correctly" in this research) a host over a period of time is a valuable capability and is applicable in fields such as network forensics and information warfare. This section will illustrate two novel fingerprinting techniques and show how they can be used to identify and re-identify hosts on the Internet in order to construct a profile for each of those hosts in dynamic IP space. Successful host tracking requires multiple fingerprinting techniques to be used in concert to create a holistic and detailed fingerprint. Four categories were defined in Section III-A,representing the different types of fingerprinting metrics required for successful host tracking. In this section we elaborate and show results based on the latency multilateration and associate fingerprinting techniques.

A. Latency Multilateration

Multilateration is a technique used to determine the location of an object by measuring the difference in distance from multiple stations with known locations by broadcasting signals at known intervals [19]. Unlike triangulation which is concerned with the measurements of absolute distance and angles, multilateration uses only timing and ranges of distance to plot multiple hyperbolic curves that intersect. This reveals a small number of potential locations, thus producing a "fix" [19]. Adapting the technique of multilateration towards network based fingerprinting involves the measurements of latency (the time it takes for a packet to reach a destination) from multiple "base stations".

For the initial investigation into latency based multilateration a proof of concept application was developed that made use of three free web based ping services, situated in geographically separate locations.

The process of latency multilateration starts with numerous ping probes that are sent from the three base stations towards a target IP address. The timings of the ping responses are recorded, outliers are removed and the average time is calculated. This produces a 3-tuple representing average time taken by each set of probes. The process of comparing these results with others to determine the likelihood of two hosts being the same is a non-trivial problem, especially with large datasets. To overcome this the 3-tuple was mapped to euclidean 3-space, representing each of the three timings as a value on an x, y and z axis.

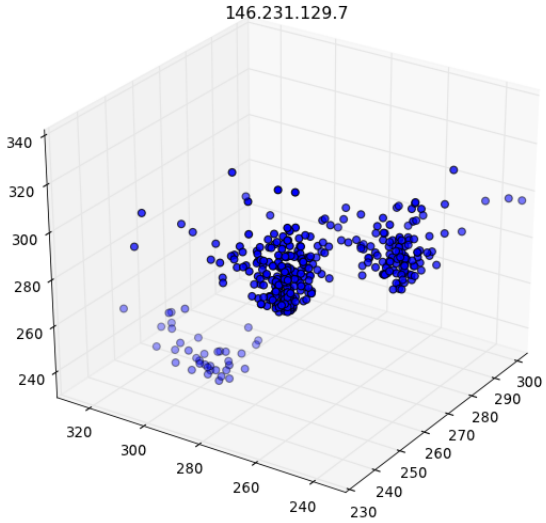

Figure 3. 3D Scatter plot of raw data without outlier removal from 10 ping scans to a single host.

This results in a series of points in space which allows for easier comparisons between hosts as the distance between two points in space is a trivial calculation as shown with Algorithm 1.

Algorithm 1 Calculate the distance between two points in space.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

A 3D scatter plot of raw time measurements is shown in Figure 3. This shows two distinct groupings of values and a small set of extreme outliers. While the dense groupings illustrate similar results over time, the presence of two separated groups also show the sporadic nature of network congestion which is the biggest concern for latency based multilateration. By applying a distance threshold to determine whether two sets of results represent the same host, we are able to create a comparable logical representation of devices that are geographically separated, thus remotely fingerprinting a physical characteristic of devices on the Internet. During our testing of latency based multilateration fingerprinting, tests against various static IP addresses over a period of time were performed. Figure 4 shows 10 scans that were run against a host with a one hour interval between each scan. The results show definite groupings of latency from each of the three base stations used. Due to the unpredictable nature of congestion and physical interferences, however, the spread of latency across the base station with the highest time difference were not ideal.

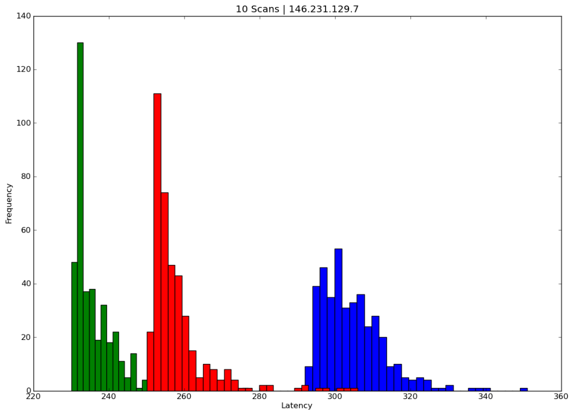

Figure 4. Latency results from three base stations targeting a specific host.

The box plots of latency results in Figure 5 show an anomaly that was encountered during testing: all three base stations revealed similar results to those shown in Figure 5. The target IP address that was being tested in this instance was under the control of these researchers, and thus it is known that while these measurements were being taken, the host experienced considerable load over a nine hour period, resulting in a visible deviation from the expected, normal latency. This result motivates that while latency based multilateration for fingerprinting holds merit, it cannot be reliably used as a single metric for fingerprinting a host.

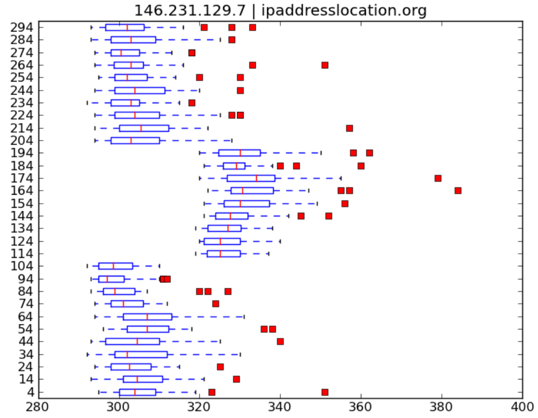

Figure 5. Timing results from a ping base station showing unreliability in latency based measurements.

For the initial investigation three free online ping tools were used to perform the testing. By making use of these publicly accessible sites it is acknowledged that the results achieved, while relatively accurate, may be improved upon by the use of dedicated ping base stations.

B. Associative Fingerprinting

As discussed in Section III-A3, we define associative fingerprinting as the identification of an affiliation of infected hosts with a given botnet. Associative fingerprinting is only applicable if we are able to identify the host as belonging to a botnet - this is done through the inspection of characteristics of a host, comparing them to known characteristics of botnets. The script created to enumerate Fast-Flux botnets continually queries a known Fast-Flux domain to record the IP addresses returned, noting the frequency with which this occurs. At the same time as harvesting IP addresses, a port scan is performed on each host. In addition to the port scan, a HTTP GET request is sent, and a record of the response from each of the hosts is kept. As mentioned in Section II-A3 these bots act as a content relay to increase the persistence of the botnet by creating thousands of hosts that hide the location of malicious content such as phishing websites. Figure 1 showed the frequency of returned IP addresses during the process of enumerating 1000 bots. This data is interesting as outliers were returned more than three times as frequently as the majority of hosts. These outlying hosts might be used by the bot herders for specific tasks such as propagating updates or hiding a second layer of Fast-Flux nameservers. These hosts might also represent the bots with the most available bandwidth and could become likely targets during a botnet take-down procedure.

The results of the HTTP GET request to port 80 found that of the 756 hosts that replied all of them shared the following two characteristics:

- The web server being used was: Apache - Nginx/0.8.34
- The last modified date in the header was set to: Sun, 11 Mar 2012 18:20:42 GMT.

It is inferred from this that the bots in the Kelihos/Hlux botnet were all configured using the same configuration script and that all of the bots were running an Nginx webserver which could act as a reverse proxy. Using these characteristics a signature to identify hosts belonging to this specific botnet has been determined. While this may not be valid in the future, it will remain useful as long as the botnet is functioning in it's present form. The bots have maintained the same HTTP responses for three weeks after initially gathering this data. Any phishing payloads present on these bots are only be accessible through a specific URL, as the bot acts as a reverse proxy,forwarding data from the “mothership”, thus HTTP GET requests to the root directory on the web server are likely to only return details regarding the initial configuration of the bot.

A basic port scan was also run against each of these hosts, however due to the time requirements of a portscan and the dynamic nature of bots connecting and disconnecting from the botnet over time only 361 of the portscans returned results. The top seven open ports are shown in Table III.

The Microsoft Windows RPC Service has, over many years, reported a multitude of vulnerabilities and it is expected that the vast majority of bots connected to a botnet would have open ports for this service as it is likely the cause of the initial compromise. From the HTTP GET responses it can be seen that all 1000 hosts were running a Nginx HTTP web server.

Port	Frequency	Description
80	361	HTTP
135	155	Windows RPC Service
139	149	Windows RPC Service
445	148	Windows RPC/SMB Service
443	56	HTTPS
593	28	Windows RPC Service
25	26	SMTP

Table III
TOP 7 TCP PORTS FROM PORTSCANS OF THE KEILIOS/HLUX BOTNET.

The port scans, however, were only able to detect 361 of these hosts. Of the 361 hosts scanned, 26 were running a mail server - these hosts were likely used to send spam emails and propagate phishing attacks. This data was obtained through the reconnaissance of the Kelihos/Hlux botnet to construct a signature of the hosts that belong to this botnet and, potentially, other similar botnets. By using this signature it is possible to differentiate between a subset of legitimate and potentially malicious hosts on the Internet. These characteristics would also assist in the fingerprinting and tracking of hosts detected over a period of time.

V. CONCLUSION

Through the application of remote host fingerprinting techniques and an adapted multisensor data fusion model this research has shown how to generate situational awareness regarding nefarious hosts on the Internet. This was achieved by identifying four categories of remote fingerprinting; software, physical, behavioral and associative. These categories provide a holistic representation of heterogeneous device characteristics. Not only do these techniques provide information regarding the physical and logical characteristics of a device but, when used as a collective, could enable unique identification and re-identification of hosts in dynamic IP space. We have created an abstract model to represent the multisensor data fusion of unsolicited network traffic, this formal method can be applied to the creation of frameworks for generating situational awareness based on unsolicited traffic. This research introduces two novel fingerprinting techniques: latency multilateration and associative fingerprinting. While it has been shown that latency-based multilateration is susceptible to network congestion and anomalies, the majority of results obtained during testing were consistent enough to warrant further investigation into this technique as a logical representation of physical locality on the Internet. Associative or affiliation based fingerprinting shows how a relationship between a host and some entity could be exploited to fingerprint that host using the relationship as a support metric in addition to other fingerprinting techniques. It is hoped that through the application of techniques outlined in this paper, sufficient situational awareness could be generated to provide insight into the climate of malicious hosts on the Internet as well as supporting the development of defensive measures to protect our networks.

ACKNOWLEDGMENT

The authors would like to thank the CSIR DPSS for funding and support of this research.

REFERENCES

[1] Michael Bailey, Evan Cooke, Farnam Jahanian, Andrew Myrick, and Sushant Sinha. Practical darknet measurement. http://www.eecs.umich.edu/fjgroup/pubs/darknet-ciss06.pdf.

[2] Richard J Barnett and Barry Irwin. Towards a taxonomy of network scanning techniques. In *Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology*, SAICSIT '08, pages 1–7, New York, NY, USA, 2008. ACM.

[3] Tim Bass. Multisensor data fusion for next generation distributed intrusion detection systems. In *In Proceedings of the IRIS National Symposium on Sensor and Data Fusion*, pages 24–27, 1999.

[4] Tim Bass. Intrusion detection systems & multisensor data fusion: Creating cyberspace situational awareness, 2000.

[5] Team CYMRU. The darknet project. Online, June 2004. http://www.cymru.com/Darknet/index.html.

[6] David Lee Hall. *Mathematical Techniques in Multisensor Data Fusion*. Artech House, Inc., Norwood, MA, USA, 1992.

[7] D.L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

[8] Uli Harder, Matt W. Johnson, Jeremy T. Bradley, and William J. Knottenbelt. Observing internet worm and virus attacks with a small network telescope. *Electron. Notes Theor. Comput. Sci.*, 151(3):47–59, June 2006.

[9] Samuel O. Hunter and Barry Irwin. Tartarus: A honeypot based malware tracking and mitigation framework. In *ISSA*. ISSA, Pretoria, South Africa, 2011.

[10] Lawrence A. Klein. *Sensor and Data Fusion Concepts and Applications*. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, WA, USA, 2nd edition, 1999.

[11] T Kohno, A Broido, and K C Claffy. Remote physical device fingerprinting, 2005.

[12] David Moore, Colleen Shannon, Douglas J. Brown, Geoffrey M. Voelker, and Stefan Savage. Inferring internet denial-of-service activity. *ACM Trans. Comput. Syst.*, 24(2):115–139, May 2006.

[13] Jose Nazario and Thorsten Holz. As the net churns: Fast-flux botnet observations. In *2008 3rd International Conference on Malicious and Unwanted Software (MALWARE)*, pages 24–31. IEEE, October 2008.

[14] Stefan Ortloff. Faq: Disabling the new hlux/kelihos botnet. Online, March 2012. http://www.securelist.com/en/blog/208193438/FAQ_Disabling_the_new _Hlux_Kelihos_Botnet.

[15] Niels Provos and Thorsten Holz. *Virtual honeypots: from botnet tracking to intrusion detection*. Addison-Wesley Professional, first edition, 2007.

[16] Colleen Shannon and David Moore. The spread of the witty worm. *IEEE Security and Privacy*, 2(4):46–50, July 2004.

[17] Matthew Smart, G. Robert Malan, and Farnam Jahanian. Defeating tcp/ip stack fingerprinting. In *Proceedings of the 9th conference on USENIX Security Symposium - Volume 9*, SSYM'00, pages 17–17, Berkeley, CA, USA, 2000. USENIX Association.

[18] Security TechCenter. Mictrosoft security bulletin ms08-067 - critical. Online, October 2008. http://technet.microsoft.com/en-us/security/bulletin/ms08-067.

[19] Alexander ter Kuile. Multilateration - mlat in action. Online, December 2009. http://www.multilateration.com/surveillance/multilateration.html.

[20] Edward L. Waltz and James Llinas. *Multisensor Data Fusion*. Artech House, Inc., Norwood, MA, USA, 1990.