Estimating the Internet Malicious Host Population while Preserving Privacy

Alif Wahid

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Computing and Information Systems
THE UNIVERSITY OF MELBOURNE

June 2013

Produced on archival quality paper.
Abstract

The Internet is a globally significant infrastructure that attracts a large number of threats posed by the population of malicious hosts within it. These threats scale with the size of the malicious host population, which makes the accurate estimation of this population an important challenge. The difficulty of this challenge is further compounded by the conflicting requirements of preserving the privacy of bystanders associated with malicious host behaviour while accurately identifying malicious host instances across the Internet.

In this thesis, we address this challenge of estimating the Internet malicious host population while preserving privacy. We begin by identifying four major research problems that have not been addressed in the literature. First is the lack of a model for host-to-address bindings. Second is the characterisation of malicious address properties. Third is the correlation of independent measurements. And fourth is the development of dynamic countermeasures. We subsequently proceed to develop novel solutions corresponding to the first three problems, while the fourth remains to be addressed in the future.

Our first contribution is the development of a probabilistic model for host-to-address bindings, which allows the number of hosts that attached to an observed address to be inferred based on privacy preserving data sets and a publicly accessible ground truth. We demonstrate the properties of this model in terms of preferential attachment and point out its primary benefit in terms of enabling the inference of host behaviour based only on address characteristics, which is a nec-
ecessary condition for privacy preservation. However, this leads to the need for an understanding of various address characteristics in order to draw reliable and robust inferences.

Our second contribution is the analysis of a large repository of intrusion alerts from globally distributed vantage points that provide access to various characteristics of malicious addresses. We find that alerted addresses are active for very short periods in the order of a few minutes and that they rarely appear more than once. We also find that there are statistically self-similar properties corresponding to these addresses in terms of non-existent temporal and spatial clusters. The main implication is that intrusion alerts contain the necessary information for use with our model of host-to-address bindings but lack sufficient robustness for reliably estimating the number of malicious hosts corresponding to an address due to the presence of spoofed and inactive sources.

Our third contribution is the combined analysis of passive measurements in the form of intrusion alerts with active measurements in the form of ping responses in order to identify those addresses that are active, attached, allocated and malicious simultaneously across two different data sets gathered independently. Our guiding hypothesis is that intrusion alerts and ping responses are different behavioural aspects of the same underlying malicious hosts. Subsequently, we apply our probabilistic model of host-to-address bindings to this intersected data set and find that more than 80% of observable addresses bind to multiple hosts, and that the distribution of malicious hosts across the IPv4 address space is highly non-uniform. This has major implications for the widespread use of blacklisting to counter the threat posed by malicious hosts, since the information used to blacklist various sources usually expires quickly.

The aforementioned overall contributions of this thesis collectively form a methodology for estimating the number of malicious hosts corresponding to an observed address while preserving privacy. We demonstrate that we can achieve
a reasonable accuracy of estimation while maintaining the privacy of all associated users. Our work is also based on openly accessible data sets and ground truth, which enables reproducibility of our results. We also demonstrate that this is broadly applicable within the Internet infrastructure that exists today.
This page intentionally left blank.
Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

__________________________________________________________
Alif Wahid, June 2013
This page intentionally left blank.
Acknowledgements

I would like to thank my supervisor, Professor Christopher Leckie, for his guidance, support and encouragement over the years. I have learned much from our many discussions about the research presented in this thesis. Above all, I am indebted to Professor Leckie for his patience in dealing with my slow grasp of everything.

I would also like to thank my collaborator, Dr. Chenfeng Zhou, for his help and valuable feedback throughout my candidature.

I am grateful to my advisory committee, comprising Professor Rao Kotagiri, Associate Professor James Bailey and Dr. Jeffrey Chan. They have provided regular feedback during progress meetings and have given plenty of encouragement at crucial stages of my candidature.

I would like to thank Professor John Heidemann from the University of Southern California for giving me access to the ping response data set. I would also like to thank Professor Paul Barford from the University of Wisconsin for providing access to DShield alerts. I am especially grateful to the people operating the CRAWDAD repository for the wealth of data sets that they make openly available to researchers all over the world.

Finally, this research was funded by an Australian Postgraduate Award, for which I am forever grateful to the Australian taxpayer.
# Contents

1 Introduction
   1.1 Internet Threat Landscape and Defences ........................................... 2
   1.2 The Need for Host Population Estimates .......................................... 5
   1.3 The Difficulties of Accurate Population Estimates .............................. 6
   1.4 The Conflict between Accuracy and Privacy ...................................... 8
   1.5 Tripartite Formalisation ................................................................. 9
   1.6 Desired Solution(s) ........................................................................... 12
   1.6.1 Critical Requirements ................................................................. 12
   1.7 Outline of Our Solution .................................................................... 13
   1.7.1 Probabilistic Model of Host-to-Address Bindings ................................ 15
   1.7.2 Characteristics of Intrusion Alerts ............................................... 15
   1.7.3 Correlation of Intrusion Alerts with Active Measurements ................. 16
   1.8 Structure of the Thesis ...................................................................... 16
   1.9 List of Publications ........................................................................... 20
   1.9.1 Journal Papers .............................................................................. 20
   1.9.2 Conference Papers ........................................................................ 20

2 A Survey of Malicious Host Estimation Methodologies ........................... 21
   2.1 Introduction ....................................................................................... 21
   2.2 Requirements of Estimation Methodologies .................................... 22
   2.2.1 Privacy Preservation ...................................................................... 22
   2.2.2 Reliable Identification ................................................................... 23
   2.2.3 Persistent Tracking ........................................................................ 24
   2.2.4 Reproducible Validation ............................................................... 25
   2.2.5 Broad Applicability ....................................................................... 26
   2.3 Review of Existing Methodologies .................................................... 26
   2.3.1 Application Layer Tracking ............................................................. 26
   2.3.2 Active Network Monitoring ............................................................. 29
   2.3.3 Passive Intrusion Detection ............................................................... 31
   2.3.4 Revised Internet Protocol ................................................................. 32
   2.4 Our Qualitative Assessment ............................................................... 34
   2.5 Open Research Problems .................................................................... 37
List of Figures

1.1 Threats arising from the Internet malicious host population ..................................... 3
1.2 The relationship between addresses, hosts and users .................................................. 10
1.3 Outline of our approach to estimating the number of malicious hosts ......................... 14

2.1 Qualitatively mapped summary of current methodologies ........................................... 36

3.1 Traditional model of a bursty traffic source ............................................................... 45
3.2 Illustration of how addresses can bind to hosts and vice versa ................................. 47
3.3 The complete space of stochastic paths for $n = 4$ appearances .............................. 49
3.4 Distribution of sample sizes ......................................................................................... 54
3.5 Absolute error of mean $\hat{p}_n$ .................................................................................. 56
3.6 Absolute error of median $\hat{\rho}_n$ .............................................................................. 57
3.7 Power-law mapping between $n$ and the mean $\hat{p}_n$ .................................................. 59
3.8 Power-law mapping between $n$ and the median $\hat{\rho}_n$ ........................................... 60
3.9 Exponential mapping between $n$ and the mean $\hat{p}_n$ ................................................ 61
3.10 Exponential mapping between $n$ and the median $\hat{\rho}_n$ ......................................... 62

4.1 Contributions of individual subscribers to the DShield repository ............................... 70
4.2 Count-frequency distribution of the full set of malicious addresses ......................... 76
4.3 Count-frequency distribution of malicious addresses ................................................... 78
4.4 Count-frequency distribution of /24 subnets .............................................................. 79
4.5 Count-frequency distribution of Horizontal addresses ................................................. 80
4.6 Count-frequency distribution of Static addresses ......................................................... 81
4.7 Count-frequency distribution of Vertical addresses ..................................................... 82
4.8 Count-frequency distribution Hybrid addresses .......................................................... 83
4.9 Distribution of address lifetimes over the IPv4 space ................................................... 86
4.10 Distribution of intrusion categories over the IPv4 space .............................................. 87
4.11 Normalised rates of new address appearances ............................................................. 88
4.12 Number of malicious addresses reported by all subscribers ...................................... 90
4.13 Number of malicious addresses reported by different subscribers ......................... 92
4.14 Number of malicious /24 subnets reported by different subscribers ......................... 93
4.15 Number of unique Horizontal addresses ..................................................................... 94
4.16 Number of unique Static addresses ............................................................................ 95
4.17 Number of unique Vertical addresses ........................................................................ 96
List of Tables

2.1 Our assessment of existing methodologies .......................... 35
4.1 Sample rows from the DShield logs ................................. 69
5.1 Chronologically ordered example of ping responses ............... 107
This page intentionally left blank.
Chapter 1
Introduction

The Internet has expanded into a globally significant infrastructure over the past three decades. It is founded on an open architecture that permits numerous scale-free networks \[1, 2, 3, 4, 5\] to interconnect with each other to provide end-to-end communication between heterogeneous devices, commonly referred to as hosts (e.g., servers, workstations, laptops, tablets, smartphones etc.). However, the Internet’s greatest strength from a growth perspective, i.e., its open architecture, is also its greatest weakness from a security and privacy perspective, i.e., the unaccountability \[6\] arising from the universally deployed IPv4 protocol \[7\] that resides at the very foundation of the Internet.

What was originally envisaged as a productive platform for its users, however, has now transformed into a less productive nuisance and a growing security threat. This is due to the activities of cyber-criminals who opportunistically exploit the vulnerabilities of the hosts connected to the Internet and convert these hosts into malicious hosts. These rogue hosts are then used for illicit activities sustained by an underground economy \[8, 9, 10, 11, 12\], ranging from the transmission of unsolicited emails en masse, through to the propagation of more malware to continually replenish the Internet malicious host population. Indeed, the Internet infrastructure itself is under constant threat from this dynamic population, which is difficult to reliably identify and track \[13, 14\].

In this thesis, we address the problem of estimating the malicious host population
while preserving privacy. We examine the difficulties of this problem and motivate its importance in relation to effective network-based countermeasures that preserve user privacy. We develop a set of critical requirements for estimating the malicious host population, and subsequently assess the existing literature to find that there is significant scope for a better trade-off in terms of the accuracy of estimation versus the privacy of users. We develop a probabilistic model of how hosts bind to addresses such that their behaviour can be reliably inferred from observing address characteristics only. This allows us to preserve the privacy of any users implicated during the course of network surveillance and measurement collection. We validate this model using a publicly accessible ground truth and apply it to real-world data sets comprising the intersection of intrusion alerts and ping responses. Overall, our contributions in this thesis comprise a novel methodology for reliably estimating the number of malicious hosts corresponding to an alerted source address in a privacy preserving manner. The remainder of this introductory chapter presents background and motivational material regarding malicious hosts, prior to discussing the research contributions and structure of the thesis.

1.1 Internet Threat Landscape and Defences

Consider the overall Internet threat landscape depicted in Figure 1.1. Various kinds of malicious hosts are continuously engaged in unwanted and illicit activities on the Internet. For example, spam emails have been a perennial nuisance for over two decades, and inaccurately classifying non-spam mail servers as spam in an aggressive effort to filter out the nuisance further compounds the problem [15, 16, 17]. Port scanning worms propagate by exploiting vulnerabilities in the networking stack of an Operating System (OS) [18, 19, 20, 21]. Overloading a legitimate web server by flooding its resources with many fake or malicious re-
1.1 Internet Threat Landscape and Defences

Each alert is the pair (Time, IP).

Figure 1.1: Threats arising from the Internet malicious host population. Notice the ambiguity arising from hosts 1 and 2 both binding to IP address 1 in this example.
quests is the basis of the well known Denial of Service (DoS) attack, which regularly plagues popular websites [22]. When an army of worm-infected hosts (commonly called a botnet) is used to launch such modes of attack, then the potent Distributed DoS (DDoS) scenario results in loss of access and service to legitimate users of legitimate websites [23, 24, 25, 26]. In recent years, smartphones have accelerated the spread of malware in ever more effective ways [27, 28].

There are numerous defensive countermeasures available to both consumers and enterprises. As shown in Figure 1.1, an Intrusion Detection System (IDS) is a widespread network based countermeasure deployed by network operators to scan traffic flows at the gateway of their networks in order to detect anomalous or suspicious behaviour. These malicious traffic flows can then be either statically blacklisted for closer scrutiny or blocked entirely. For example, among the most popular IDSs currently in deployment are Snort [29] and Bro [30]. The former relies on signature analysis of reassembled packet streams to look for various malware binaries, while the latter relies on an event-driven paradigm to detect unauthorised traffic flows into and out of a network. In addition, smaller scale firewalls that function in a similar way are also widely deployed by consumers in residential broadband routers [31]. Network intrusions can include promiscuous behaviour, such as port-scanning or when we observe unusual patterns of accesses between groups of users who do not normally interact [32].

In addition to network based countermeasures, many organisations also deploy host based countermeasures. Among the most popular host based countermeasures is anti-virus software designed to scan for known worm and Trojan binaries executing in the background. A generally reliable way of eliminating many existing malicious hosts is the prompt application of all security patches and updates released by the various OS and anti-virus vendors [31]. These patches and updates generally fix exploitable vulnerabilities in both the OS and the anti-virus software, as well as helping to recover a potentially compromised host. How-
ever, for various administrative and economic reasons, security updates are not always promptly applied. For example, downloading large patches en masse can easily consume expensive bandwidth, which often leads to the scheduling of updates less frequently. This creates a lag between the outbreak of a worm and the corresponding security patch. Furthermore, these patches cannot counter novel threats (commonly called zero-day attacks) and are reactive by definition. Firewalls and IDSs, on the other hand, provide a more proactive method of monitoring and countering zero-day attacks as well as existing malware propagating on the Internet [33]. For these countermeasures to be effective, we need to be able to accurately monitor the malicious host population that is the source of these attacks.

1.2 The Need for Host Population Estimates

Estimating the Internet malicious host population is essential for two important reasons. First, the scale of the threat posed by a specific population of malicious hosts strongly correlates with the size of that population [12]. The implication is that an individual malicious host is seldom the only adversary that computer networks need to defend against. Instead, it is the dynamic population of malicious hosts on the Internet that consumes the majority of effort regarding network monitoring and countermeasures [34]. Second, specific countermeasures, such as the proactive blacklisting of entire network domains, have been proposed [35]. A key factor in deciding whether to blacklist a domain is the size of the malicious host population that is resident in that domain. The open question is how to accurately estimate this population size. Hence, this thesis focuses on the major challenges surrounding the estimation of the size of the population of Internet malicious hosts while preserving the privacy of users.
1.3 The Difficulties of Accurate Population Estimates

The malicious host population is heterogeneous in composition and highly dynamic in terms of individual host life-cycles [36, 37, 25, 33]. The combination of network and host based countermeasures leads to a situation whereby individual malicious hosts undergo cyclical transitions from being vulnerable to being infected and eventually patched [38, 25, 26]. This creates churn and volatility in the malicious host population such that identifying and tracking individual hosts becomes a real challenge. The primary reason for this difficulty is the limited range of data that one can generally access in order to estimate the malicious host population. Internet-wide traffic flow measurements can be indexed by the conventional 5-tuple format comprising source address, destination address, source port, destination port and protocol fields extracted from packet headers at any point of monitoring or surveillance [39, 40, 41], such as the IDS and firewall gateways shown in Figure 1.1.

Consider the seemingly straightforward task of identifying hosts on the Internet based on any 5-tuple network measurement. Even though at any given instant there is usually a tight binding between an Internet host and its corresponding IP address, this close association invariably expires quickly due to the many ways that IP addresses are shared and recycled. For instance, in a typical Local Area Network (LAN) the Dynamic Host Configuration Protocol (DHCP) [42] is used for leasing IP addresses (both static and dynamic) to client hosts for fixed time intervals in the order of a few hours to a few days [43]. In residential broadband connections the Remote Authentication Dial-In User Service (RADIUS) [44] protocol is generally used for dynamically allocating addresses for a few hours at a time prior to recycling [45].

Consequently, over the course of much longer intervals of many weeks or months, it becomes a challenge to resolve the address aliases when using source addresses to identify and track individual malicious hosts (note that existing alias
1.3 The Difficulties of Accurate Population Estimates

Resolution techniques [46, 47, 48, 49, 50, 51] are targeted toward the problem of resolving the numerous addresses of multi-homed routers. In addition, private LANs are generally connected to the global Internet via fixed gateways that decouple the internal address pool from the globally visible address space using a Network Address Translator (NAT) [52, 53, 54, 55]. The effect of this is that any external measurement of the traffic originating from a NAT will observe one address corresponding to an unknown number of hosts (some of which can easily be malicious). This also occurs for hosts that are hidden behind web proxies, shared firewalls and the Domain Name System (DNS) [56] (DNS hiding is exploited by Fast Flux Service Networks to efficiently propagate malware [57, 26]).

Statistics reported in the literature demonstrate the unreliability of IP addresses as identifiers of unique hosts. For instance, Xie et al. [58] found that over 30% of addresses that accessed a large webmail service were only static for less than three days. Heidemann et al. [59] observed that the median time for visible Internet hosts to be continuously bound to their respective addresses was just 81 minutes. Maier et al. [45] observed that half of the addresses used by a European Internet Service Provider (ISP) in a large suburb were assigned to at least two different residential broadband lines in the space of 24 hours, while 1-5% were assigned to 10 different lines during the same period. Rajab et al. [13] and Casado et al. [60] have estimated that anywhere from 19% to over 60% of malicious hosts reside behind NAT devices respectively. In addition, Kreibich et al. [61] observed that 90% of user web-sessions originated from behind NAT domains while Maier et al. [62] found that more than 50% of residential broadband lines contain multiple hosts hidden behind such domains permanently. This volatility of the binding between a host and its corresponding IP address motivates the need for an accurate and reliable method of estimating the number of malicious hosts corresponding to observed addresses. Having an accurate estimate of the malicious host population enables more effective and targeted network based countermeasures, such
as dynamic and proactive blacklisting of suspicious addresses [63, 15, 35, 64].

1.4 The Conflict between Accuracy and Privacy

The aforementioned technical difficulties of estimating the malicious host population are due to a lack of measurement data. In principle, it may be possible to alleviate this challenge by gaining access to precise measurements that fully identify hosts and their corresponding addresses at all times. For instance, the Media Access Controller (MAC) that connects a host to an Ethernet LAN has a globally unique address that binds tightly to the host over long periods of time. This information is indeed used by DHCP for assigning IP addresses to individual hosts [42]. However, in practice, there are legal and ethical barriers that typically do not permit access to such accurate data due to the requirement of preserving user privacy and their right to anonymity [65, 66]. This creates a conflict between the accuracy of estimation and the privacy of users that must be resolved by any applicable solution.

In order to illustrate the tensions between the requirements for accuracy and privacy, we highlight two recent studies reported in the literature. Triukose et al. [67] found that geolocating smartphone users via the IP addresses of their cellular connections can lead to errors in excess of 100 km for 70% of the users, along with severe uncertainty when it comes to actually differentiating between users due to the recycling and translation of IP addresses by cellular network operators. In this case, user privacy and anonymity are largely preserved (i.e., their geographic location) at the expense of inaccurate inferences drawn about the population of smartphone hosts (i.e., identifying and tracking the total number of unique devices in a particular region). On the other hand, Yen et al. [68] found that client browser details, source addresses, cookies and webmail logins can all be used to accurately identify and track hosts at the expense of user privacy. They
further report the ability to reliably track users’ online activities (which is a potential violation of their right to anonymity) even when cookies are deleted and browser caches are purged.

1.5 Tripartite Formalisation

The problem of estimating the malicious host population while preserving privacy is inherently a data driven activity. Different kinds of estimates can be derived based on different kinds of measurements. As mentioned earlier, network measurements can typically be indexed by the 5-tuple format, which is a useful conception for analysing end-to-end traffic flows. The main disadvantage of this abstraction, however, is that it does not take into consideration privacy preservation in terms of the interconnecting relationship between hosts, addresses and users. So in this section, we present a formalisation of the relationship between them in order to concisely frame the development of a set of critical requirements for our estimation problem in the next section.

The three primary entities in our conception are hosts, addresses and users. A host is an instance of an OS with its own networking stack, which may have been compromised or infected by malware (i.e., we are concentrating on malicious hosts). This separates the physical hardware from the software, such that multiple virtual hosts may run on one machine. An address is the 32-bit IPv4 source identifier that can be assigned to a host when it is connected to a network for routing purposes. This may be globally unique for connecting to the Internet or locally unique for connecting to an intranet, or both. Addresses can be recycled across multiple hosts and reused by the same host with a period of detachment in between. Any permutation of those two possibilities is also valid (i.e., address aliasing). A user is a person whose identity (e.g., username, password) can be linked to a unique pair (host, address) at any given instant. We denote the set of
Figure 1.2: The relationship between addresses, hosts and users in our tripartite formalisation. We can always observe addresses and the task is to infer their interconnections to hosts. However, in the process of doing so, users may become implicated and that raises legal barriers with respect to accessing sensitive data.

hosts, the set of addresses and the set of users through the lower case letters $h$, $a$ and $u$ respectively, along with a subscript character for enumeration of their individual elements (e.g., $h_i$ is the $i^{th}$ unique host, $a_j$ is the $j^{th}$ unique address and $u_k$ is the $k^{th}$ unique user).

Definition 1.1. An **instantaneous binding** at time $t$ for a monitored network is the undirected tripartite graph $B_t = (V_{ht}, V_{at}, V_{ut}, E_{(h-a)_t}, E_{(h-u)_t})$, where $V_{ht}$ is the set of vertices corresponding to hosts, $V_{at}$ is the set of vertices corresponding to addresses, $V_{ut}$ is the set of vertices corresponding to users, $E_{(h-a)_t}$ is the set of edges connecting $V_{ht}$ to $V_{at}$, and $E_{(h-u)_t}$ is the set of edges connecting $V_{ht}$ to $V_{ut}$.

As a consequence of Definition 1.1, addresses can be connected to hosts, and hosts can be connected to users, such that there are no direct connections between users and addresses. This reflects the separation of users from addresses, which provides a key insight about privacy preservation, i.e., if a unique pair $(h_i, a_j)_t$ is available then it is capable of reliably identifying the user $u_k$ at time $t$ when there is a strong association between users and hosts. This is further illustrated in Figure 1.2.

Definition 1.2. A **dynamic binding** is the $N$-tuple $B_T = (B_{t_1}, B_{t_2}, ..., B_{t_N})$, which corresponds to a measurement interval comprising $N$ discrete time instances, $T = (t_1, t_2, ..., t_N)$,
and captures the temporal volatility of the edges that connect addresses to hosts and hosts to users.

**Definition 1.3.** A dynamic host-to-address binding is the undirected bipartite graph $B_{(h-a)_T} = (V_{h_T}, V_{a_T}, E_{(h-a)_T})$, which is a strict subset of $B_T$ for the corresponding measurement interval $T$ and only contains the union of interconnections between addresses and hosts such that $V_{h_T} = \bigcup_{t \in T} V_{h_t}$, $V_{a_T} = \bigcup_{t \in T} V_{a_t}$, and $E_{(h-a)_T} = \bigcup_{t \in T} E_{(h-a)_t}$.

**Definition 1.4.** The population of hosts at time $t$ is the cardinality $|V_{h_t}|$ and the population of hosts over the interval $T$ is the cardinality $|V_{h_T}|$.

In the same manner as Definition 1.4, the population of users and the population of addresses can be defined. However, estimating those populations is not the focus of this thesis, since addresses are self-evident in measurements while users are to be hidden in order to protect their privacy. This means that the population of hosts is the intermediary that connects users to addresses at any time $t \in T$. Therefore, an exact determination of the population of hosts will always violate user privacy by impinging on anonymity.

**Definition 1.5.** The act of privacy preservation equates to the deletion of a sufficient number of edges from $E_{(h-u)_t}$ such that a user cannot be linked to a unique pair $(h_i, a_j)$, that reveals his or her identity with a threshold probability of $p > 0$ at any time $t \in T$.

As a consequence of Definition 1.5, the only guaranteed method of privacy preservation is the deletion of all edges from $E_{(h-a)_t}$, in order to force the equality $B_T = B_{(h-a)_T}$ and reduce the tripartite conception to a bipartite one. The practical implication is that the use of any sensitive data (that links users to their corresponding hosts) in order to estimate the malicious host population may be prohibited by privacy laws in certain jurisdictions, including Australia [65, 66]. Consequently, a suitable trade-off between accuracy and privacy is needed in this context.
1.6 Desired Solution(s)

The life-cycle of individual malicious hosts makes the collective population \( |V_{hT}| \) a difficult quantity to estimate \([69, 26]\). Furthermore, the requirement of privacy preservation means that the necessary data are either inaccessible or must not be used during analysis. Therefore, both the possibility and the utility of estimating \( |V_{hT}| \) is uncertain. As pointed out by Rajab et al. \([14]\), the notion of cardinality when it comes to a botnet is volatile given the multitude of ways that the botnet can morph and evolve. Instead, it is more useful to know the number of hosts corresponding to an alerted address since that is what blacklisting is fundamentally reliant on. Moreover, accurately inferring the distribution of hosts across the global IPv4 address space can reveal regional malicious clusters, as well as removing the inconvenience of spoofed and inactive addresses that frequently raise false alarms \([32, 63, 70]\). This also opens up the possibility of identifying previously blacklisted addresses that are no longer malicious due to the inherent malware life-cycle, i.e., the patching of infected hosts. Hence, our proposed methodology in this thesis addresses the specific problem of estimating the number of malicious hosts corresponding to an alerted address.

1.6.1 Critical Requirements

Any solution to this problem of estimating the number of malicious hosts corresponding to an alerted address needs to satisfy a set of critical requirements that we propose in terms of striking a trade-off between the accuracy of estimation and the privacy of users. These requirements are listed below in descending order of priority.

1. Privacy Preservation
2. Reliable Identification
3. Persistent Tracking
4. Reproducible Validation
5. Broad Applicability

First, privacy preservation must be given the highest priority since access to network measurements is generally contingent on this being satisfied. Second, reliable identification of malicious hosts is required based on the limited information available about observed addresses. Third, persistent tracking of any identified hosts is essential so that the volatility and aliasing effects of dynamic addressing are overcome. Fourth, we propose that any method of estimation should be validated using publicly accessible ground truth such that others can reproduce the inferences made. Finally, any method of estimating the number of malicious hosts corresponding to an alerted address needs to be applicable on the current Internet infrastructure as opposed to being a thought experiment about future architecture(s).

These five critical requirements provide a consistent way of assessing and guiding the search for a solution. We elaborate further and justify their completeness in Chapter 2 along with assessing the existing literature to search for better trade-offs than what is currently possible. What follows next is an outline of our solution to this problem that comprises the remaining chapters of the thesis.

1.7 Outline of Our Solution

The overall work-flow of our estimation methodology is illustrated in Figure 1.3. There are three major parts to our approach that are discussed individually as follows.
Figure 1.3: Outline of our approach to estimating the number of malicious hosts corresponding to individual source addresses while satisfying the five critical requirements listed in the upper left-hand quadrant.
1.7 Outline of Our Solution

1.7.1 Probabilistic Model of Host-to-Address Bindings

In the upper right-hand quadrant of Figure 1.3, the first step in our proposed methodology is the development of a probabilistic model of host-to-address bindings. This model allows the calculation of conditional probabilities for the number of hosts corresponding to an address given: a) the number of times the address appeared in the measurements, and b) its latent probability of binding to a new host in each appearance. Our aim here is to design a statistical framework that can enable inferences to be made from the limited amount of accessible data. However, reliable identification and tracking of malicious hosts can only be carried out by applying this model to real-world measurements that provide only the characteristics of addresses that bind to hidden malicious hosts. As a result, this step alone does not satisfy all of the critical requirements listed in the upper left-hand quadrant of Figure 1.3.

1.7.2 Characteristics of Intrusion Alerts

In the lower right-hand quadrant of Figure 1.3, we add a second step in our proposed methodology to establish the characteristics of intrusion alerts that aid in identifying malicious behaviour without violating the privacy of hosts and users concerned. In this case, we exploit the utility of passively generated alerts from Collaborative Intrusion Detection Systems (CIDS) like the DShield.org repository in accessing multiple global vantage points for monitoring malicious behaviour that manifests as port-scanning. However, while we find that intrusion alerts do provide the necessary basis for applying our model of host-to-address bindings, there remains sufficient uncertainty in the observed address characteristics such that reliably estimating the number of malicious hosts corresponding to an address is still not fully possible from these two steps combined.
1.7.3 Correlation of Intrusion Alerts with Active Measurements

In the lower left-hand quadrant of Figure 1.3, we add the third and final step in our proposed methodology, whereby we correlate intrusion alerts with active measurements in the form of ping responses, such that various uncertainties can be sufficiently resolved. For example, spoofed and inactive addresses can be effectively filtered out by repeatedly pinging them over many days and weeks, as demonstrated in [59]. The independence between active ping responses and passive intrusion alerts provides adequate confidence that those addresses which both generate intrusion alerts and reply to ping queries are indeed bound to malicious hosts over long periods of time. Hence, such a correlated data set can be used in conjunction with our model of host-to-address bindings for reliably estimating the number of malicious hosts corresponding to an alerted address without violating the privacy of users who may become implicated in the process of gathering the necessary measurements. Furthermore, these three steps collectively satisfy all of the critical requirements.

1.8 Structure of the Thesis

The remaining chapters of this thesis focus on the various challenges discussed earlier along with our proposed solution. The synopsis of each chapter is as follows.

Chapter 2 - A Survey of Malicious Host Estimation Methodologies carries out a comprehensive survey of the relevant literature in the context of the conflicting requirements of accuracy and privacy. We utilise the aforementioned tripartite formalisation of the estimation problem involving hosts, addresses and users to classify the existing approaches available in the literature within a single framework. Subsequently, we critically examine the current methodologies and qualitatively map them onto the orthogonal dimensions of accuracy and privacy.
1.8 Structure of the Thesis

The main conclusion here is that a better trade-off can be reached most effectively via a hybrid approach based on the correlation of independent data sets comprising both passive and active measurements. The main contribution of this chapter is the following:

- We assess the existing methods of estimating the malicious host population in terms of accuracy and privacy, and highlight the open research challenges that are addressed in subsequent chapters of the thesis.

Chapter 3 - A Probabilistic Model of Host-to-Address Bindings aims to develop the desired trade-off between accuracy and privacy identified in Chapter 2. By utilising the phenomenon of preferential attachment, we show that the number of hosts corresponding to an address is accurately predicted by the number of times that an address appears in a series of alternating ON and OFF intervals. We validate this method using a four month trace of dynamic address allocations in a campus wireless network. In so doing, we demonstrate the practical significance and utility of such a privacy preserving method for estimating the number of hosts corresponding to a dynamic address. The main contributions of this chapter are the following:

- We develop a parsimonious stochastic model of host-to-address bindings based on the phenomenon of preferential attachment, which allows the calculation of conditional probabilities for the number of hosts corresponding to an address given the number of times the address appeared and its latent probability of binding to a new host in each appearance.
- We propose approximations to this model for efficiently handling large data sets without compromising accuracy.
- We validate this model using anonymised DHCP logs collected from a campus wireless network over the course of four months.

The conference paper directly resulting from the work in this chapter is [72].

Chapter 4 - Self-Similar Characteristics of Intrusion Alerts analyses logs
from the DShield.org repository of globally distributed IDSs in order to test the hypothesis that network intrusion attempts exhibit self-similar characteristics, such that their suitability for use with the proposed model of host-to-address bindings in Chapter 3 can be established. Three forms of evidence in favour of the self-similar hypothesis are presented. First, the lifetime of malicious hosts conform to a power-law, such that the overwhelming majority are short-lived while a few are long-lived. Second, the distribution of malicious hosts across the IPv4 address space is broadly identical regardless of different categories of lifetimes. Third, there is a scale invariant diurnal cycle with long-range dependence in the number of unique hosts observed per unit-time. The main contributions of this chapter are the following:

- We analyse intrusion alerts and find self-similar characteristics that make them suitable for use with the proposed model of host-to-address bindings.
- We point out the practical implications of self-similar characteristics that affect static blacklisting of malicious addresses, in which the volatility of IPv4 addresses is not accounted for.

The journal paper directly resulting from the work in this chapter is [73]. Some of the work from this chapter also appeared in an earlier conference paper [74].

Chapter 5 - Correlation of Intrusion Alerts with Active Measurements investigates the underlying trends of visible Internet hosts that were randomly selected from the global pool of IPv4 addresses and actively probed using ping measurements every 11 minutes for two weeks [59]. The existence of self-similar characteristics is demonstrated via the power-law relationship that regulates the appearance of Internet hosts as alternating sequences of ON and OFF intervals, along with the diurnal cycle that exhibits long-range dependence. Such characteristics make this data set applicable for use with the proposed model of host-to-address bindings. We then combine the contributions of Chapters 3 and 4 into the proposed complete solution by applying the probabilistic model to the inter-
section of the ping response and the intrusion alert data sets. The subsequent finding is that more than 80% of the addresses identified and tracked as malicious actually bind to multiple hosts. Furthermore, the distribution of malicious hosts across the IPv4 address space is highly non-uniform and is suggestive of geographical clusters of focused malicious activity on the Internet. The major contributions of this chapter are the following:

- We examine actively probed ping responses from 1% of the globally visible Internet address space and find self-similar characteristics that make them suitable for correlation with intrusion alerts and the proposed model of host-to-address bindings.
- We demonstrate the benefits of correlating intrusion alerts and ping responses that were independently collected. In particular, the correlation of these two independent data sets provides the means for reliably identifying and tracking malicious hosts on a per address basis.
- We apply the proposed model of host-to-address binding and infer that more than 80% of malicious addresses that are reliably identified actually bind to multiple hosts.
- We discover that the distribution of the number of malicious hosts per address is highly non-uniform across the IPv4 address space, which is suggestive of geographical clusters of malicious hosts (e.g., localised botnets).

The journal paper directly resulting from the work in this chapter is [75].

Chapter 6 - Conclusions and Future Research summarises the contributions and findings of the thesis, as well as discussing possible directions for future research. In particular, we discuss the scope of extending the proposed model of host-to-address binding based on the use of non-parametric Bayesian techniques.
1.9 List of Publications

1.9.1 Journal Papers


1.9.2 Conference Papers


Chapter 2
A Survey of Malicious Host Estimation Methodologies

2.1 Introduction

The main benefit of the tripartite formalisation presented in Chapter 1 is that it concisely captures the objective of privacy preservation, which is a core requirement of solving this problem of estimating the malicious host population. In addition, the concept of a dynamic host-to-address binding highlights the temporal volatility that must be resolved in order to accurately identify and track malicious hosts. So in this chapter, we utilise the aforementioned formalisation and definitions in order to survey the existing methods of estimating the malicious host population. We discuss the requirements of estimation methodologies with a view towards making a better trade-off between accuracy and privacy than what is currently possible. We also identify the scope for making this trade-off and discuss the open research problems that need to be addressed in the process. Note that this chapter provides a comprehensive background survey, while more specific material related to all the contributions of this thesis appear in the subsequent chapters that follow.

The remainder of this chapter is organised as follows. Section 2.2 elaborates further on the set of critical requirements for estimation methodologies outlined previously in Chapter 1. Subsequently, Section 2.3 surveys the literature accord-
ing to these criteria. Our qualitative assessment of the literature is presented in Section 2.4. A discussion of the open research problems based on our review and assessment of the literature can be found in Section 2.5. Finally, we summarise and conclude this chapter in Section 2.6.

2.2 Requirements of Estimation Methodologies

In order to carry out a comprehensive survey of the existing literature on estimating the malicious host population, we formulate a set of criteria by which to assess any given methodology as follows.

2.2.1 Privacy Preservation

The importance of privacy preservation arises from the legally guaranteed right to anonymity in various jurisdictions around the world. For example, in Australia this is legislated at both the federal and state levels of government. The Federal Privacy Act of 1988 explicitly defines “Information Privacy Principles” that severely restrict the collection, storage and analysis of information that can, at any time, uniquely identify a person [66]. Furthermore, the Victorian Information Privacy Act of 2000 extends the reach of these principles to network measurements [65]. Equivalent policies and corresponding statutes exist in the United States as well, e.g., refer to Mayer and Mitchell [76] for a discussion of the policies and technologies that are geared towards preventing users from being tracked by various organisations.

Further examples that bring into focus the inherent privacy violations of accurately inferring users’ online activities as well as the economic incentives for businesses can be found in the recent study by Friedland et al. [77]. Also refer to [78] regarding an insightful discussion of the need to ignore personal attribu-
tion while countering multi-stage attacks, where the authors highlight the need for privacy preservation as opposed to the accurate identification of criminals as being the primary objective.

The practical implication of this requirement is that any method of estimating the malicious host population needs to prioritise privacy preservation first in order to rank highly in our assessment. As per Definition 1.5, this requires that a set of users associated with a set of hosts must be explicitly disconnected so as to keep their anonymity intact.

### 2.2.2 Reliable Identification

Conventional analysis of network security data sets use IPv4 source addresses as static identifiers of corresponding hosts. This would reduce the estimation problem to simply counting the number of unique addresses observed, that is, $\|V_{ht}\| = \|V_{at}\| \forall t \in T$. However, there is ample evidence in the literature \cite{79,45,59,58,43} that demonstrates the limitations of assuming a static binding between addresses and hosts. This is because the inequality $\|V_{ht}\| \neq \|V_{at}\|$ holds over the course of a long measurement window (e.g., several weeks) even though the instantaneous equality $\|V_{ht}\| = \|V_{at}\|$ may hold only temporarily.

One potentially reliable method is host authentication based on user credentials, which corresponds to the edge set $E_{(h-u)}$, connecting hosts and users along with the concomitant edge set $E_{(h-a)}$, connecting hosts and addresses at any time $t \in T$. In this manner, being able to link a user to a unique pair $(h_i, a_j)_t$ means that one may reasonably assume the host $h_i$ is tightly coupled to the address $a_j$ at time $t$ such that address $a_j$ is a reliable identifier of host $h_i$ for the purpose of inferring the number of unique hosts (i.e., estimating the malicious host population). However, removing a sufficient number of edges from $E_{(h-u)}$ for privacy preservation would necessarily invalidate this assumption due to temporal volatility.
and address aliasing. Therefore, the critical question that we ask when surveying existing methods is: \textit{how is user privacy preserved while utilising a tight coupling between a given host and address pair?} 

\subsection{2.2.3 Persistent Tracking}

Provided that a method exists to reliably identify hosts at any given instant, then the challenge becomes one of \textit{persistently tracking} individual hosts over the course of a long measurement interval. Two successive static bindings can easily differ in the way that the corresponding addresses connect to hosts, and vice versa. That is to say it is possible that \( B_{t_x} \cap B_{t_y} = \emptyset \) for \( \{t_x, t_y\} \in T \). So the union of the edge sets over the course of a measurement interval will necessarily incorporate temporal volatility of the bindings between addresses and hosts. Therefore, the task of inferring a dynamic host-to-address binding \( B_{(h-a)_T} \) requires a \textit{method of tracking an individual host across its bindings to different addresses while preserving privacy.}

This is a non-trivial challenge because hosts can be \textit{single-homed} (i.e., attached to one address at any one time) or \textit{multi-homed} (i.e., attached to multiple addresses at any one time). Moreover, hosts may \textit{evolve} over the course of a measurement interval from being single-homed to multi-homed and vice versa. They may also undergo the classic birth-death process, whereby they are only visible temporarily with the persistent arrival of new hosts (i.e., births) and the disappearance of old hosts (i.e., deaths). Our formalisation accommodates all these possibilities due to the versatility of the graph abstraction, since \( B_{(h-a)_T} \) comprises a sequence of instantaneous static bipartite graphs connecting hosts and addresses over the course of the measurement interval \( T \). Thus the requirement is to identify \textit{changes} in the bipartite graph while compiling \( B_{(h-a)_T} \).

The architecture of the Internet creates more challenges for this requirement of
2.2 Requirements of Estimation Methodologies

Persistently tracking hosts. NATs and proxies can cause multiple distinct hosts to appear as one host over the course of a measurement interval, even if it is possible to reliably identify a host at any given instant (as illustrated in Figure 1.1). The question then becomes how to disambiguate the illusion of one host when there are multiple hosts in reality? This again requires violating user privacy, in principle, to resolve hosts easily.

Assuming that one user is attached to one host behind a NAT or proxy, then theoretically it is possible to disentangle the host that appears as a NAT or proxy into its hidden constituent hosts by enumerating the associated user set at the application layer [68]. For instance, this could be done by tracking webmail logins of users situated behind a proxy server [80]. In other words, for any given $h_i$ if the corresponding $|u| > 1$ then it is likely to be made up of multiple hidden hosts. However, privacy preservation would make this impossible since the edge sets connecting hosts to users will be nullified. In addition, malicious hosts can reside in mobile devices and/or be associated with different users. Consequently, it is by no means guaranteed that even tracking associated users will resolve this kind of ambiguity. One notable exception to this rule is the observation by Bellovin [81], whereby it is possible to exploit the incrementing ID field of IPv4 packet headers as a way of demarcating specific hosts that exhibit such a signature even if they are situated behind a NAT.

2.2.4 Reproducible Validation

Estimation methods need to be reproducible and accessible to the broader network security research community for the scientific method to be upheld [82, 83, 84]. It is foreseeable, in theory at least, that privacy is preserved quite easily by not sharing the underlying ground truth used to validate a proposed method of estimating the malicious host population. We would consider that to be a major
2.2.5 Broad Applicability

Any potential method that requires a fundamental reinvention of the Internet’s architecture in order to easily handle the previous requirements cannot be an attractive proposition due to the economic costs arising from legacy incompatibility. Thus it is critical that any technique be broadly applicable in terms of both backwards compatibility and usability on the current Internet infrastructure distributed around the world [34, 87]. This requirement is in contrast to the one articulated in [88], where the focus is on innovation. We are sceptical about the possibility of architectural innovations finding broad applicability in the near future given the history of the Internet and its stability over many years.

2.3 Review of Existing Methodologies

This section surveys the current literature with respect to the requirements developed in the previous section. We classify existing methods into four categories and examine each category in detail as follows.

2.3.1 Application Layer Tracking

The core feature of this category of methods is the compilation of information and events from the application layer with concurrently observed IP addresses from the network layer in order to satisfy the technical requirements of reliably identifying and persistently tracking malicious hosts. We have identified three
approaches from the literature that belong to this category of estimation methodology as listed below.

1. **Relying on the binding between users and hosts in order to infer a tight coupling between hosts and addresses at any given instant.** Xie et al. [80] and Yen et al. [68] utilise this approach in their method of identifying and tracking malicious hosts.

2. **The use of application level Internet infrastructure in order to identify malicious address domains.** Bilge et al. [89] rely on this method to identify and track domains that hide malicious hosts behind very few visible addresses.

3. **Dedicated applications for measuring host-to-address bindings at specific instants with the aid of volunteer participants.** Kreibich et al. [61] and Casado et al. [90] have deployed web applications for the purpose of gathering measurements from the Internet edge that rely on volunteer cooperation and participation to get around privacy restrictions.

In the first approach, Xie et al. [80] propose a system called “HostTracker” that records the IP addresses from which users access their webmail accounts. In conjunction with assuming a tight coupling between users and hosts at the time of login, their system is capable of inferring a dynamic host-to-address binding as per Definition 1.3. They report more than 90% of the hosts being tracked correctly over the course of a month using proprietary ground truth collected during the application of security updates. This same principle was subsequently extended by Pitsillidis et al. [91] and Yen et al. [68] in order to include other application layer events and information, such as browser cookies, with a similar outcome with regards to identifying and tracking hosts.

The accuracy of this approach is excellent, and several subtle technical difficulties were resolved by Xie et al. [80]. However, the need for a proprietary ground truth that is unavailable for reproducible validation is a major drawback. Yen et al. [68] also suffer from the same disadvantage due to their use of propri-
etary HotMail and Bing data sets provided by Microsoft. Moreover, the privacy preservation aspect is unconvincingly handled in this approach due to the highly sensitive data being gathered without the explicit consent of the users. This restricts the application of this method to covert and proprietary environments outside the purview of the public Internet in which we are interested.

In the second approach, Bilge et al. [89] propose a system called “EXPOSURE” that classifies DNS traffic in order to detect malicious domains by using 15 underlying traffic “features” that are reported to predict malicious activity quite well. They categorise these features into four groups: a) Time-based, b) DNS answer-based, c) Time-To-Live (TTL) value-based, and d) Domain name-based. This allows them to build a supervised classifier that processes a DNS traffic trace offline in order to infer malicious domains and generate an estimate of the population for each domain. They report a worst-case false-positive rate of 20%, with the majority of the feature-based classification having around 10% error.

The specificity of EXPOSURE to DNS is the main reason for its poor accuracy compared to the approach taken by HostTracker. However, the privacy preservation is significantly better since malicious behaviour is identified in the form of aggregated domains without requiring any association between users and hosts in terms of our tripartite formalisation. The volume of DNS information also puts this approach at a disadvantage as filtering must be applied to manage the offline workload [89]. The validation method is likely not to be reproducible either due to the supervised classification scheme that relies on features of DNS traffic that are prone to concept drift over time (i.e., tomorrow’s attackers can easily workaround any statistical profile discovered during the training phase today). Nevertheless the reliance on current DNS infrastructure makes this a broadly applicable method.

The third approach utilises custom applications that volunteer participants can execute on their machines. Kreibich et al. [61] developed a platform called
“Netalyzr”, which uses a digitally signed Java applet that must be executed by each participant in order to provide sensitive information such as the current IP address, operating system details, local network topology and so on. Casado et al. [90] used the same approach in terms of a Java applet to gather detailed information about the participants’ host environments. Even though these are not explicitly targeted for estimating the malicious host population, they can be used to do so with very little modification in the worst case.

The final approach differs from the previous two in terms of the instantaneous reliability of the data gathered. Netalyzr can execute sophisticated tests on the client to decisively answer questions about the host environment (e.g., is it behind a NAT?). However, due to concept drift, there is little prospect of accurately tracking hosts over long intervals subsequent to a volunteer’s one-time participation. However, the privacy concerns are better handled due to the digitally signed applets being voluntarily executed by the participants. Kreibich et al. [61] report more than 130,000 executions in the space of several months as an indication of Netalyzr’s popularity and broad applicability. Unfortunately, the anonymised data gathered are not publicly accessible, which prohibits reproducible validation. Also note that a recent derivative of Netalyzr is “Fathom” [92].

2.3.2 Active Network Monitoring

As the overwhelming majority of malicious hosts are herded into botnets, a fruitful approach to estimating their population has been active network monitoring [8, 69, 93]. The core of this approach is to engage malicious hosts unwittingly into revealing their attack methods and packet payloads. This is primarily achieved by the deployment of distributed honeynets, whereby unallocated or unused address ranges are populated with hosts that act as honeypots in order to deliberately engage and setup connections with any suspicious packets destined to these hon-
eynets [93]. Even in the absence of honeypots, it is possible to actively store all packet payloads that wander into these darknets. Typically, such active network monitoring is effective for well known botnets and their corresponding propagation vectors [33].

Rajab et al. [38] proposed a multi-faceted honeynet architecture in order to investigate various botnets. They tracked Internet Relay Chat (IRC) based botnet command and control servers over the course of three months and found 192 botnets of various sizes ranging from a few hundred to several thousand malicious hosts. Tellingly, they discovered that more than 10% of the 800,000 DNS domains they monitored contained malicious hosts. In terms of privacy preservation, this approach satisfies our basic requirements due to the fact that it is restricted to worms that propagate via known vectors (e.g., port scanning), and that these vectors are not reliable at inferring potential user activities associated with the corresponding malicious hosts. Any port scanning behaviour is typically an algorithmic signature of malicious hosts as opposed to a behavioural pattern of the innocent users of those hosts. Further advances have been made in automating data analysis [94] and managing honeynet infrastructure [95] in recent years.

Reliably identifying malicious hosts and persistently tracking them using honeynets is generally limited to existing botnets that rely on IRC based command and control servers as well as propagating via port scanning. Consequently, even though the accuracy of estimation is excellent for well known botnets in this approach, new threats and zero-day attacks are not typically well handled [33]. Weaver [96] provides insightful evidence of the modelling and data gathering necessary to partly overcome this drawback at the cost of lagging in response to new malware such as the Conficker-C worm. There is also the new problem of actually setting up distributed darknets on the Internet due to exhaustion of the global IPv4 address pool [97], which impacts their broad applicability criteria. The data gathered by this approach are typically not publicly accessible either,
which unfortunately limits the scope for reproducible validation [93].

2.3.3 Passive Intrusion Detection

A widespread defensive measure against port-scanning is the use of an IDS that passively monitors traffic flows at the gateway of a network connected to the Internet and generates alerts of external intrusion attempts by logging the corresponding flow attributes of the traffic (e.g., time, source IP, destination port, etc.). Such passively collected information provides valuable evidence for administrative and forensic actions like threat analysis [20], detecting botnets plus malware infections [98, 99], and collaboration with other networks for a global view of malicious activity on the Internet [70].

There are two main IDSs found on the Internet today, Snort [29] and Bro [30]. Snort is predominantly reliant on signature analysis, whereby classic attack methods like buffer overflow payloads can be detected by looking for specific patterns embedded within the reassembled packet stream at the gateway to a private network. Bro, on the other hand, is event driven and allows the specification of various behavioural patterns to classify traffic. Both of these provide immense scope for customisation towards the specific needs of any given private network, which is the main reason for their popularity with enterprise network managers and operators.

Individual IDSs lack the oversight necessary to detect coordinated attacks and new malware propagating on the Internet. As a result, Collaborative Intrusion Detection Systems (CIDSs) have become popular in recent years, e.g., DShield.org [71]. Zhou et al. [100] contains a comprehensive survey of various coordinated attack methods and CIDSs. Apart from the global oversight provided by a CIDS, they also provide sufficient privacy in the form of anonymised data sharing since any collaboration would not be possible otherwise. There is also significant pre-
dictive power in the alerts generated by a CIDS due to the aggregation of independent vantage points on the Internet [63]. However, their major disadvantages reside in reliably identifying and persistently tracking malicious hosts due to the fact that dynamic addressing issues discussed in Chapter 1 are not resolved in any reliable way. Also, intrusion detection in general is a passive measurement technique, which is susceptible to address spoofing [101].

2.3.4 Revised Internet Protocol

This category of methods involves clean-slate proposals for re-architecting the network layer in order to eliminate the current unaccountability of addresses. They propose inherently reliable methods of host identification that work across long time-scales on the order of weeks to months. The most relevant new architecture currently available in the literature is called the Accountable Internet Protocol (AIP) [6]. Central to this proposal is the idea of certified addresses, whereby both domains and hosts can attest to the validity of the addresses to which they are attached. In order to achieve this, AIP discards the current Classless Inter-Domain Routing (CIDR) prefixes with a proposed two-level hierarchy. The upper level corresponds to independent Accountability Domains (ADs) that are equivalent to today’s Autonomous Systems (ASs), while the lower level corresponds to individual hosts that are associated with a globally unique Endpoint Identifier (EID). Thus the AIP address is structured as AD:EID when attached to a host. The domain to which a host belongs maintains public keys that certify its uniqueness and accountability, which are then used to securely hash an identifier for assignment as an EID to a host. Therefore, the essence of AIP is to rely on cryptographically secure signatures as the fundamental means of identifying hosts and domains. This core idea has been extended by two other proposals, namely “IPa+” [102] and “DHT-MAP” [103].
From the perspective of estimating the malicious host population, AIP and the relevant extensions offer excellent solutions to the technical challenges we have outlined. Hosts can be reliably identified and persistently tracked over the course of a long measurement interval without directly violating user privacy. The creation of ADs provides a layer of indirection that separates the users from being directly identifiable, even though the associated hosts can be identified as being an accountable sub-domain of their upper level AD. However, being a clean-slate solution means that this is unlikely to be deployed and will most likely remain a thought experiment. Such two-level addressing schemes have implicit scalability and traffic engineering limitations (although not the focus of this thesis) that make them unpalatable for deployment on the Internet. Consequently, it is unfortunately a non-starter as far as a solution is required for estimating the malicious host population that has broad applicability.

Another proposal in this category is the well known Internet Protocol version 6 (IPv6) that contains 128-bit source and destination addresses in order to exponentially expand the exhausted 32-bit IPv4 address pool [104]. Such an immense address space practically removes the need for any hierarchy that exists currently for hiding private networks from the public Internet (i.e., NATs and proxies). Even though IPv6 does not mandate that an address be bound permanently to its host, it is one possibility that can easily emerge with severe privacy implications as to how anonymity can be guaranteed in that instance. Moreover, the long standing reluctance for enterprise and commercial network operators to deploy IPv6 en masse is testament to the fact that such clean-slate proposals for re-architecting the Internet have an uncertain future. Refer to the recent measurements by Zander et al. [105] and Dhamdhere et al. [106] regarding the small fraction of global Internet hosts that are IPv6 capable and the correspondingly small percentage of observed traffic classified as IPv6.
2.4 Our Qualitative Assessment

In Table 2.1 we have summarised our assessment regarding the strengths and weaknesses of the four categories of estimation methodologies surveyed in the previous section. We have adopted two simple scales. The initial three requirements of privacy preservation, reliable identification and persistent tracking are judged in terms of a ternary scale comprising poor, adequate and excellent. The remaining two requirements of reproducible validation and broad applicability are judged in terms of a binary scale comprising yes and no. Admittedly, we have adopted coarse-grained scales in assessing the literature but that is because the diversity of methodologies do not lend themselves to a consistently finer-grained judgement. We can also graphically map these existing categories of estimation methodologies to the orthogonal dimensions of accuracy and privacy, as illustrated in Figure 2.1.

In our assessment, revised Internet protocols like AIP provide the best potential trade-off in this space. However, their clean-slate nature makes them virtually unusable on the backwards compatible Internet that we have inherited today. Out of the remaining three categories of methodologies, application layer tracking provides the best accuracy in our assessment, although this comes at the expense of violating privacy preservation. This also makes them unsuitable given that both active network monitoring and passive intrusion detection are generally better at preserving privacy albeit at the expense of accuracy. It is also worth noting that there are considerably more instances of passive intrusion detection and active network monitoring than application layer tracking in use today, which provides significant scope for a hybrid approach to achieve a better trade-off as indicated in Figure 2.1.
Table 2.1: Our assessment of existing methodologies across the four categories surveyed. The requirements of reproducible validation and broad applicability are characterised in terms of a binary scale comprising Yes and No, while the remaining three requirements are judged in terms of a ternary scale comprising Poor, Adequate and Excellent.
Figure 2.1: Qualitatively mapped summary of current methodologies for estimating the malicious host population in terms of accuracy versus privacy.
2.5 Open Research Problems

Based on the preceding survey and assessment of existing methodologies, the following research questions arise in search of a better trade-off between accuracy and privacy.

1. Is there an underlying model of host-to-address bindings that can be utilised in order to reliably identify and persistently track malicious hosts? Given that the volatility of IP addresses is typically tamed with privacy violating measurements (e.g., [80, 45]), there is a need to develop empirically reliable models that can describe the way hosts bind to addresses. In this way, accurate inferences might be drawn regarding the population of hosts corresponding to an observed address while preserving privacy. To the best of our knowledge, there is no such model that is explicitly derived to encapsulate dynamic host-to-address bindings.

2. How can one validate any model of host-to-address bindings without violating privacy? This is an issue revolving around access to ground truth data sets that are essential for reaching reliable conclusions. In particular, what level of anonymisation is an adequate safeguard while keeping intact the inherent statistical characteristics of the ground truth? The general lack of publicly accessible ground truths is perhaps one reason for the absence of a host-to-address binding model.

3. What kind of hybrid measurements are necessary to accurately estimate the malicious host population while preserving privacy? In particular, what are the statistical characteristics of the available data that directly impact the scope for finding a better trade-off? What are the implications for predictability based on any potential trend found in the data? The underlying properties of measurements gathered via active network monitoring and passive intrusion detection have not been correlated to produce a possibly better trade-off as far as we know in the relevant literature.
4. What are the practically compelling applications of achieving the identified trade-off when it comes to countering the malicious host population? The requirement here is to explore the implications of existing countermeasures for defeating the threats posed by the Internet malicious host population. In particular, how is blacklisting affected by the volatility of IP addresses and the requirement of privacy preservation? Is it possible to devise an effective dynamic policy of blacklisting whereby addresses automatically transition between being blacklisted and being unlisted based on up-to-date relevant information?

The pursuit of answers to the aforementioned research questions can be grouped into four separate endeavours, which we discuss as follows.

2.5.1 Underlying Model of Host-to-Address Bindings

Our tripartite formalisation of the estimation of malicious hosts suggests the need for an underlying model of host-to-address bindings that can handle the dual technical requirements of reliable identification and persistent tracking in addition to the privacy preservation constraint. Among all of the methods that we have surveyed, no such model was explicitly proposed. It has been shown that the process of address allocations is not entirely random, whether at the level of private LANs or the public Internet [97, 79]. Consequently, there is scope for taking advantage of any inherent behavioural patterns comprising the bindings between hosts and addresses in order to empirically model and test their predictive power. Incidentally, modelling the transient nature of network addresses was highlighted by Partridge recently among his list of unanswered research questions in data communications (#16 on the list of 40 problems) [87].

It is also worth noting that the lack of a publicly accessible ground truth remains a problem since that severely restricts any concerted effort at reproducible
validation of potential models. We are only aware of a four month DHCP trace in the public CRAWDAD repository that was collected at a campus wireless network during 2005 \[107\]. One possible remedy could be the mediated access procedure proposed by Mittal et al. \[108\], whereby a data owner carries out various analysis on behalf of a remote data analyst and only the results are made available instead of the uncensored data. The disadvantage is the bottleneck in authentication and communication required between the data owner and the data analyst, which stops this from scaling to widespread adoption in the research community. Refer to \[109\] for an in-depth review of further challenges associated with making ground truth data sets publicly accessible.

2.5.2 Characteristics of Malicious Addresses

It is necessary to characterise the spatial and temporal properties of malicious addresses in order to identify the measurable quantities that can be used in a model of host-to-address bindings. Due to privacy preservation, one can typically observe \textit{address characteristics} and not the underlying \textit{host characteristics}. However, it is an open question as to how much of the malicious host behaviour can be reliably inferred from the observed address characteristics without implicating any associated set of users. The relevant literature contains insightful findings about the observed structure of IP addresses, both across space \[110, 79\] and over time \[111, 101\]. It would be particularly useful to investigate how such measurable address characteristics correlate with their corresponding host population. In other words, despite the transient nature of addresses attached to hosts, is there any implicit predictive power in the observed address characteristics? The only relevant study in that regard that we are aware of is \[43\], which looked at the characteristics of leased addresses over DHCP and how different allocation policies perform in terms of the utilisation of a limited pool of private addresses by mobile hosts.
2.5.3 Correlation of Independent Measurements

This may provide considerable advantages in terms of drawing reliable inferences as well as striking the identified trade-off illustrated in Figure 2.1. As Wilcox et al. [16] demonstrated by correlating address characteristics with spam blacklists, the practice of statically blacklisting spamming servers is flawed when the dynamic nature of address volatility and aliasing means that legitimate servers will be rendered inaccessible very quickly. Another example is that of Ding et al. [32], whereby intrusion alerts were combined with flow-level measurements to draw out the promiscuous behaviour of malicious addresses that were engaged in multiple communities of non-malicious addresses.

We find that there is scope for applying this tactic to the estimation of the malicious host population, given that in so doing, accuracy can only increase while the necessary anonymisation can satisfy privacy preservation. However, without a valid model of host-to-address bindings it is not clear how any correlated results could be interpreted and actioned appropriately [84]. In other words, independent measurements might be better understood as reflecting different facets of the same underlying phenomenon caused by malicious hosts. For example, aggregated intrusion alerts [71] could be correlated with global ping surveys [59] in order to identify a subset of addresses that were bound to active malicious hosts. The rationale being that neither measurement on its own will reveal such a collection of addresses but only their intersection may provide the necessary reliability in terms of mitigating spoofed, unbound and non-malicious addresses.

2.5.4 Dynamic Countermeasures

The potential for outcomes from the previous three endeavours naturally leads to the challenge of how they would be used to continuously counter the threats posed by the Internet malicious host population. From a network management
and operations perspective, a model of host-to-address bindings based on the observed characteristics of addresses could be applied to the intersection of independent measurements for reaching reliable and accurate conclusions about how to automatically classify addresses as malicious, and when to update their corresponding classifications based on the arrival of new information. The requirement of privacy preservation also adds to the challenges of such dynamic blacklisting, as detailed in [64]. Another avenue is that of proactive blacklisting of suspected domains in order to counter emerging malware outbreaks before they reach critical mass [35].

Ultimately, it is the external addresses intruding into a private network that garner the most attention as far as security and monitoring are concerned, since the uncertainties regarding their volatility and aliasing are greatest (presuming that a network operator can always resolve internal address volatility and aliasing). As a result, we envisage that future efforts targeted at designing dynamic countermeasures will require input from underlying models of host-to-address bindings, characteristics of observed malicious addresses and correlation of independent measurements as vital steps in developing consistent and hybrid methodologies.

2.6 Conclusions

In this chapter, we developed a set of requirements for assessing the existing literature for various estimation methodologies. The outcome of our survey and assessment is that there is significant scope for a better trade-off in terms of the accuracy of estimation versus the privacy of users. We identified the need to develop a model of host-to-address bindings as the major open problem that needs to be solved before a potentially better trade-off can be achieved. To that end, we also identified the possibility of correlating independent measurements as a
potential means to solving this trade off. The overarching objective of these re-
search problems is to develop more accurate methods of estimating the malicious
host population in subnetworks, so that targeted countermeasures can be more
effectively deployed. In the remainder of this thesis, we begin by addressing the
modelling challenge in Chapter 3. We then investigate the characteristics of ma-
lusive addresses and all that can be reliably inferred from them in Chapter 4.
Subsequently, we demonstrate the validity of correlating independent data sets
in Chapter 5. Finally, we discuss dynamic countermeasures as possible future
extensions in Chapter 6.
Chapter 3
A Probabilistic Model of Host-to-Address Bindings

3.1 Introduction

The rationale for a probabilistic model of host-to-address bindings is twofold. First, in order to find the desired trade-off between accuracy and privacy, we require a statistical framework that can enable inferences to be made from the limited amount of accessible data. Second, any estimation task involving large volumes of data is far more applicable and testable when founded on a parsimonious model of the underlying phenomenon. Hence, our objective in this chapter is to develop a behavioural model of the way that hosts bind to addresses that is based only on the measurable characteristics of addresses, not hosts. Our main contributions are as follows.

- We develop a probabilistic model of host-to-address bindings based on the notion of preferential attachment. This model allows the calculation of conditional probabilities for the number of hosts corresponding to an address, given: a) the number of times the address appeared, and b) its latent probability of binding to a new host in each appearance.
- We derive and test two computationally tractable approximations of this model using a large data set of dynamic host-to-address bindings from a campus wireless network that comes with anonymised ground truth re-
garding the exact number of hosts corresponding to an observed address. However, this ground truth is purely for validation and our model is only dependent on the two stated parameters related to the observed addresses (as opposed to parameters related to the hidden hosts, which potentially violate user anonymity).

The rest of this chapter is structured as follows. We outline some background theory in Section 3.2 to concisely frame and contextualise the subsequent formulation of our model in Section 3.3. We develop approximations of the proposed model in Section 3.4 and validate them using an anonymised ground truth in Section 3.5. Finally, we discuss some related work in Section 3.6 and conclude this chapter in Section 3.7.

### 3.2 Theoretical Background

The need for parsimony when fitting a model to a set of observations is important given that it is easy to fall into the trap of overfitting in any modelling exercise [112]. The risk of overfitting is exacerbated by the lack of publicly accessible ground truth data sets for validation due to privacy restrictions. As a result, we build on a model from the theory of bursty network traffic, in which it is possible to model a traffic source as a time-series of alternating ON and OFF periods that are independent and identically distributed (i.i.d.) [113, 114]. This is illustrated in Figure 3.1. Statistical self-similarity and long-range dependence can emerge if the ON or the OFF periods exhibit heavy-tailed distributions, denoted $t_{ON}$ and $t_{OFF}$ respectively, with infinite variance [115, 116]. From the perspective of network measurements, this manifests as bursty traffic from any given address at any given time-scale that can be partitioned into discrete sequences of contiguous ON and OFF periods [116, 117]. The ON or the OFF periods are typically distributed according to a power-law probability density that is parameterised
3.2 Theoretical Background

I.I.D. series of ON and OFF periods for an address

Corresponding state machine representation with appearance counter 'n'

Figure 3.1: Traditional model of a bursty traffic source (i.e., an address) as an alternating series of independent ON and OFF periods. The time-series representation (top) can subsequently be translated into a state machine representation (bottom) that counts the number of discrete “appearances” by an address.

by a scaling exponent and a tail threshold [118, 119].

It is important to emphasise that the ON and OFF periods are defined in terms of a probability distribution and not precise temporal partitions. The implication is that it is not possible to exactly partition a time-series of packet arrivals from an address into an alternating series of ON and OFF periods due to the many independent possibilities inherent in deciding where an ON period ends and its adjacent OFF period begins (or vice versa). However, it is possible to compute the higher order statistical properties of a time-series (e.g., the Hurst parameter [116]) that derive the power-law parameters for best resembling the ON and OFF periods in a distributional sense. This is a long standing result in the traffic modelling literature [113, 114, 115, 116, 117, 118] and its proof is beyond the scope of this thesis. What is essential for our modelling purposes is the fact that one of
these periods is distributed according to a power-law with infinite variance \[116\].

The insight that we rely upon in using this model is the concept of preferential attachment \[1\], whereby an address is solely attached to one host during any given ON period, and subsequently, the state transition (as per Figure 3.1) from one ON period to the next provides an opportunity for it to attach to either a brand new host or another host that it attached to in the past. As a result, the discrete appearances of an address become a measurable quantity that allows us to hypothesise about the number of hosts corresponding to that address over the course of an observation window \(T\).

As per Definition 1.3 from Chapter 1, a dynamic host-to-address binding is the bipartite graph \(B_{(h-a)_T} = (V_h, V_a, E_{(h-a)_T})\). Given that we are only concerned with hosts and addresses in this chapter, we drop the excess subscripts from the formal notation such that a host-to-address binding is simply \(B = (V_h, V_a, E_{h-a})\) and refers to the same Definition 1.3. On the top right-hand corner of Figure 3.2, we depict the host-to-address binding corresponding to an example time-series on the top left-hand side.

In the absence of any data that would easily give away the exact binding \(B\), we can only attempt an estimation on probabilistic grounds by using the measurable number of appearances, denoted \(n\), for any given address \(a\). This uncertainty is illustrated in the bottom half of Figure 3.2 where the labelled host attachments have been removed such that one cannot distinguish precisely which host bound to which address for how long and exactly when.

### 3.3 Formulation of Our Model

In building a parametric model of host-to-address bindings, our objective is to minimise the number of measurable and hidden parameters from the outset. We can only measure the number of appearances by an address, denoted \(n\), from any
3.3 Formulation of Our Model

Figure 3.2: Illustration of how addresses can bind to hosts and vice versa.

HYPOTHETICAL = EXACT host-to-address binding available WITHOUT privacy

REAL-WORLD = IMPOSSIBLE to infer exactly which host bound to which address WITH privacy

Figure 3.2: Illustration of how addresses can bind to hosts and vice versa.
network related data set. Then the question becomes: how do we express the notion of preferential attachment using the measurable number of appearances by an address?

Our solution is to add one more hidden parameter that is to be learnt from a ground truth data set. So we define the conditional probability $P(h_a \leq k|n, p_n)$, where $h_a$ is the number of hosts corresponding to an address $a \in V_a, k \in \{1, 2, 3, \ldots n\}$ and $p_n$ is a latent parameter that specifies the finite independent probability with which an address $a \in V_a$ may bind to a new host at its $n^{th}$ appearance (conversely, $\bar{p}_n = 1 - p_n$ specifies the probability that the address may bind to a previously attached host at the $n^{th}$ appearance).

The conditional probability $P(h_a \leq k|n, p_n)$ encompasses the deviation of the binding $B$ from the non-aliasing expectation of static address binding, whereby $h_a = 1 \forall a \in V_a$ and $|V_a| = |V_h|$. Moreover, the monotonically increasing nature of an address’s appearance count $n$ implies that those addresses that have already attached to multiple hosts in the past are preferred by new hosts that may appear in the future. We now formally define the estimation problem stemming from this model as follows.

**Definition 3.1.** Given a series of alternating ON and OFF periods corresponding to each address $a \in V_a$, then the conditionally expected number of hosts per address, $E[h|n, p_n]$, for the underlying binding $B$ is:

$$E[h|n, p_n] = \frac{1}{|V_a|} \sum_{a \in V_a} \sum_{k \in n} k \cdot P(h_a = k|n, p_n) \quad (3.1)$$

The main assumption of our model so far is that of independence: both between addresses and between appearances. This means that any given address traces out a random path from a full binary tree that is $n$ levels deep. Figure 3.3 demonstrates one such address with $n = 4$ appearances, which has a corresponding parameter vector $p_n = \{p_1, p_2, p_3, p_4\}$ specifying its binding probabilities for new hosts at the $n^{th}$ appearance, and conversely, the parameter vector
3.3 Formulation of Our Model

\[ p_{n} = \{\bar{p}_{1}, \bar{p}_{2}, \bar{p}_{3}, \bar{p}_{4}\} \] specifying the binding probabilities for old hosts at the \( n^{th} \) appearance.

As is evident, the first appearance can only mean one host and so we can expect \( p_{1} \) to be 1 with its complement being 0. However, the subsequent appearances must diverge in a binary manner depending on whether a new host or an old host is attached to the address. Thus, after the fourth appearance we have eight separate possibilities that arise due to the assumption of independence between appearances (note that in some cases the same number of hosts is inferred via mutually independent paths). Therefore, the space of possibilities for the number of hosts corresponding to an address grows exponentially with the number of times that the address appears during the measurement interval.
3.4 Approximating the Binding Probabilities

Over the course of many weeks and months it would be perfectly normal for an address to appear hundreds of times, which means that the time complexity of computing Equation 3.1 is in $O(|V_a| \cdot 2^n)$. This necessitates finding an approximation in order to make the computation tractable for large data sets with thousands of independent addresses and their corresponding appearance counts spanning more than two orders of magnitude.

Our approach to approximating the exponentially growing space of possibilities for one address is to assume that the binding probabilities are constants for any given number of appearances. That is to say that we reduce the full binary tree comprising a vector of $n$ independent binding probabilities to the Binomial distribution with one binding probability that covers all the paths of the tree for a corresponding $n$. In other words, we have a forest of $n$ fixed depth binary trees that are individually indexed to Binomial distributions for computing the conditional probability of the number of hosts corresponding to any given address. As a result, the illustrated scenario of Figure 3.3 and all other cases can be approximated by

$$P(h = k|n, p_n) = \binom{n}{k} (p_n)^k (1 - p_n)^{n-k}$$

for any fixed $p_n \in (0, 1]$ and $k \in \{1, 2, 3, ..., n\}$. However, the main issue that remains unsolved is how to estimate this fixed $p_n$ for any given $n$ with minimal loss of information. That is to say that we require a mapping function $f : \mathbb{N} \rightarrow (0, 1]$ that exploits any inherent relationship (e.g., correlation) between $n$ and $p_n$ that may exist.

We propose two variants of this mapping function for empirically testing whether there is indeed a relationship between $n$ and $p_n$. The first is a power-law model of the form

$$p_n = c_1 \cdot n^{-\alpha}$$
where $\alpha$ is a free parameter in $(0, 1]$ and $c_1$ is a scaling constant in $(0, 1]$. This means that we can conjecture a correlation between $p_n$ and $n$ of the following form

$$\ln(p_n) = \ln(c_1) - \alpha \ln(n)$$

(3.4)

which can be tested via linear regression in the log-log domain assuming log-normally distributed errors (which is not unreasonable in this case [120]). The second variant is an exponential model of the form

$$p_n = c_2 \cdot e^{-\lambda n}$$

(3.5)

where $\lambda$ is a strictly positive free parameter and $c_2$ is a strictly positive scaling constant. This allows us to conjecture a correlation between $p_n$ and $n$ of the following form

$$\ln(p_n) = \ln(c_2) - \lambda n$$

(3.6)

which can also be tested via linear regression in the log-linear domain assuming log-normally distributed errors.

These two variants have different implications although both contain one free parameter, which is essential since any more would be a case of over-fitting an already explicit approximation. The power-law variant allows a slower decay than the exponential form. This means that as an address appears more often, we inherently expect it to have an increasing number of hosts due to preferential attachment, but the rate of this increase is what differs between the two variants. This is apparent when we substitute the two ways of mapping $p_n$ while calculating the mean of the Binomial distribution as follows.

$$\mu_\alpha = n \cdot p_n = n \cdot c_1 \cdot n^{-\alpha}$$

(3.7)

$$\mu_\lambda = n \cdot p_n = n \cdot c_2 \cdot e^{-\lambda n}$$

(3.8)
Equations (3.7) and (3.8) also provide a consistency check by ensuring that the mapping functions can be readily substituted into the Binomial distribution in place of \( p_n \) without leading to any singularities. The equation for the variance of the Binomial distribution also satisfies the substitution of these functions.

Note that we chose the method of linear regression since it is a simple and efficient technique that has been used in relevant studies [3, 110] to estimate the correlation between two sequences. In this method, one minimises the sum of squared distances between a reference sequence \( X = \{ x_1, x_2, \ldots, x_N \} \), with corresponding mean \( \mu_x \), and the resulting best fit linear sequence \( Y = \{ y_1, y_2, \ldots, y_N \} \), with corresponding mean \( \mu_y \). The test for statistical significance assumes that errors are distributed normally in the log-log domain and we use a \( p \)-value of 0.01. We also measure the goodness-of-fit of the linear regression using the sample coefficient of determination, denoted \( r^2 \), defined in Equation 3.9. The values of \( r^2 \) range continuously from 0 to 1: 1 indicates perfect correlation between \( X \) and \( Y \), while 0 indicates no correlation.

\[
r^2 = \left( \frac{\sum_{i=1}^{N} (X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N} (X_i - \mu_x)^2 \sqrt{\sum_{i=1}^{N} (Y_i - \mu_y)^2}}} \right)^2
\]

(3.9)

### 3.5 Validation using DHCP Logs

#### 3.5.1 Data Format

In order to test our proposed approximations and the validity of the underlying model, we used a set of sanitised DHCP logs [107] collected at a campus wireless network over the course of nearly four months from midnight 19th of April 2005 to midnight 8th of August 2005. Each row in this data set contains the following three fields.

1. UTC time-stamp
2. 32-bit private IP address with the /16 prefix 192.168.0.0
3. 48-bit anonymised MAC address

These fields collectively specify the instant when an independent IP address was “leased” by an anonymous host (as identified by its MAC address). Each unique time-stamp corresponding to a recurring IP address directly specifies a new appearance. Moreover, each new appearance can attach either to a new host or an old host as determined via the corresponding set of MAC addresses. Therefore, the total number of appearances by an IP address is simply the cardinality of its corresponding set of time-stamps.

It is worth examining the question of how reliable are MAC addresses? They can certainly be spoofed if administrator privilege is accessible on certain systems. However, we have verified that this data set was collected in a controlled environment such that MAC address spoofing was extremely unlikely [107]. This is because the DHCP server was configured to only allocate internal IP addresses to those MAC addresses that were registered with the network beforehand (i.e., the typical practice of managed Ethernet switching). As a result, we can safely assume that a MAC address uniquely identifies a host at all times without any kind of aliasing or spoofing in this particular ground truth.

This data set is the only publicly accessible ground truth of which we are aware, where public accessibility is a critical requirement for the research purposes of this thesis. Furthermore, the long measurement interval of nearly four months along with the observation of more than seven thousand distinct hosts in this data set provides sufficient scale and the necessary diversity for reliable validation of the proposed model of host-to-address bindings. It naturally captures various modes of address recycling through frequent attachment and detachment of hosts, as well as the scenarios involving one user with multiple mobile devices.
Figure 3.4: Distribution of sample sizes, as measured by the number of independent addresses for each value of $n$. Note that in our analysis we excluded all the samples where there were less than three corresponding independent addresses.

### 3.5.2 Computing Estimates

We found a total of 6,554 independent IP addresses in these logs that were leased multiple times by over 7,000 independent hosts. This resulted in an over subscription rate of around 10% during the course of the measurement interval, which is consistent with other studies of DHCP logs [43]. We filtered the raw data down to 6,033 addresses by ensuring that for any given number of appearances there were at least three corresponding independent IP addresses in order to form a sufficiently large sample and increase the statistical power of our hypothesis testing. The distribution of sample sizes is illustrated in Figure 3.4.
We estimated the empirical parameter $\hat{p}_n$ from the data by taking the ratio of the true number of hosts $h_a$ attached to an IP address $a$ and its corresponding appearance count $n_a$, which is to say that $\hat{p}_n = \frac{h_a}{n_a}$ for any given address $a$. This is generally an unbiased estimator of the Binomial distribution’s mean (which is $\hat{p}_n$ in this case). Subsequently, for each sample comprising a set of independent addresses $V_a$, all having the same appearance count $n_a$, we used both the sample mean and the median statistics corresponding to $\hat{p}_n$ in order to be robust in our hypothesis testing. This also requires the characterisation of the difference between the true number of hosts $h_a$ attached to an address $a$ and the estimated number of hosts $\hat{h}_a$ attached to that same address based on the Binomial assumption of Equation 3.2. The following equation specifies this absolute mean error $\Delta_{\text{avg}}$ (measured in the number of hosts per address) for the sample corresponding to any fixed appearance count $n_a$

$$\Delta_{\text{avg}} = \left[ \frac{1}{|V_a|} \sum_{a \in V_a} |h_a - \hat{h}_a| \right]$$ (3.10)

where $\hat{h}_a = n_a \cdot \hat{p}_n$. Subsequently, two different distributions result for the absolute error of estimating the number of hosts attached to an address based on both the mean and the median $\hat{p}_n$ learned from the DHCP ground truth. These are plotted in Figures 3.5 and 3.6 respectively.

Evidently, the absolute error of estimating the number of hosts corresponding to an address based on the Binomial assumption of Equation 3.2 is negligible for all appearance counts of $n \leq 10$. Furthermore, this error is small in the region $10 < n \leq 100$. However, it is significant for appearance counts of $n > 100$. Such a distribution is due to the skewed sample sizes that we presented earlier in Figure 3.4. All samples in the region $1 \leq n \leq 100$ are sufficiently large such that the effect size of our estimates is significant. But this significance is reduced for small sample sizes in the region $n > 100$. This means that care is required when
Figure 3.5: The absolute error of estimating the number of hosts attached to an address based on the mean \( \hat{p}_n \) learned from the DHCP ground truth.
3.5 Validation using DHCP Logs

Figure 3.6: The absolute error of estimating the number of hosts attached to an address based on the median $\hat{p}_n$ learned from the DHCP ground truth.
applying our model to any real-world data sets to ensure that the vast majority of the samples are distributed in the reliable region of \( 1 \leq n \leq 100 \). It is also worth noting that the mean and the median \( \hat{p}_n \) yield virtually indistinguishable results, which provides a necessary consistency check in order to ensure robustness against outliers.

### 3.5.3 Regression Results

The results of linear regressions involving the power-law and the exponential mapping functions between the empirical \( \hat{p}_n \) and the observed \( n \) are plotted in Figures 3.7, 3.8, 3.9 and 3.10. In all four figures, the best-fit linear regressions were statistically significant at the 0.01 level assuming normally distributed errors in the logarithmic domain. This means that we cannot reject either of the two mapping variants even though, upon closer scrutiny, the power-law function is a better fit than the exponential function according to the coefficient of determination \( r^2 \). In the best case, our approximation can explain more than 70\% of the variance between the data and the proposed model (Figure 3.7) while in the worst case this drops to around 50\%. The power-law mapping appears to fit very accurately in the region \( 1 \leq n \leq 10 \) where we have consistently large samples in the vicinity of one hundred addresses. They both appear to fit the tail of the data equally well for \( n > 100 \), which is where the error is most apparent due to small samples of less than ten independent addresses (as evident from Figures 3.5 and 3.6).

The existence of two clear, overt correlations between \( p_n \) and \( n \) is the major finding of our evaluation, which validates the approach we took in approximating our proposed model. As a result, we can use either of these correlations to compute a conditional probability distribution for the number of hosts \( P(h = k|n, p_n) \) by substituting back into Equation (3.2). The validity of the rela-
Figure 3.7: Power-law mapping between $n$ and the mean $\hat{p}_n$ of the corresponding sample.
$n = \text{Number of Appearances by an Address}$

$\hat{p}_{\text{med}} = \text{Median Binding Probability of an Address}$

Data (312 points from 6033 addresses)

Power-law model ($\alpha = 0.61, r^2 = 0.61$)

Figure 3.8: Power-law mapping between $n$ and the median $\hat{p}_n$ of the corresponding sample.
3.5 Validation using DHCP Logs

\[ n = \text{Number of Appearances by an Address} \]

\[ \hat{p}_{\text{avg}} = \text{Mean Binding Probability of an Address} \]

Data (312 points from 6033 addresses)

Exponential model \((\lambda = 0.0019, r^2 = 0.55)\)

Figure 3.9: Exponential mapping between \(n\) and the mean \(\hat{p}_n\) of the corresponding sample.
Figure 3.10: Exponential mapping between $n$ and the median $\hat{p}_n$ of the corresponding sample.
tionship between $n$ and $p_n$ means that the expected value of this distribution can be a reliable estimate for the typical number of hosts per address in real-world data sets that either contain the number of appearances directly or allow it to be estimated in an unbiased manner.

3.6 Related Work

Our proposed discriminative model is a variant of the two-state Markov Modulated Poisson Process (MMPP) that is well documented in the literature (refer to [121] for an extensive survey). Our focus on the fractal phenomenon is closer to a special case of the MMPP known as a Pareto Modulated Poisson Process (PMPP) [122]. These processes are primarily concerned with the clustering of bursty traffic into alternating series of ON and OFF periods.

In the case of MMPP, these periods are exponentially distributed with a finite variance, whereas PMPP allows for power-law distributions with infinite variance in order to capture the emergent statistical self-similarity. We, on the other hand, are concerned with counting the preferential attachment of a latent variable that can take place during state transitions in both of these processes regardless of how the sojourn times in each state are actually distributed. In that respect, our model can be thought of as a special case of the classic Hidden Markov Model (HMM) [123].

Furthermore, ours is in the discriminative category while HMMs belong in the generative category. What this means is that the former is more restrictive due to the computation of a conditional posterior distribution $P(h|n, p_n)$, as opposed to a full joint distribution $P(h, n, p_n)$. Note that a discriminative model is more appropriate in our application context since a generative model by definition does not permit privacy protection given its complete coverage over all the variables and parameters (both hidden and visible).
Various applications of preferential attachment in the context of computer networks have been well documented in the literature over the years. Refer to [120] for an extensive survey. In the context of dynamic address allocation and volatility, there is a growing body of topical measurements that highlight the effects of preferential attachment between hosts and addresses. Refer to [61, 79, 45, 59, 58] for various detailed measurements of address volatility and the resulting uncertainty.

Dominant among such uncertainties is the presence of NATs that hide private networks (often large enterprise and campus networks, although nowadays more commonly residential broadband routers [62]). To that end, Bellovin [81] proposes a method of counting hosts hidden behind NATs by using the “Identification” field of IPv4 packet headers, which is sometimes implemented as an incrementing counter. However, this method is sensitive to OS variations and is likely to only work with a specific NAT as opposed to all NATs in general.

Another recent use of preferential attachment for estimating the host population corresponding to a dynamic address is HostTracker [80], which makes use of application-level user events, e.g., webmail logins, in order to infer a tight binding between hosts and addresses for sufficiently long enough that the bipartite graph mapping hosts to addresses can be derived accurately. However, this was not the trade-off that we were after since the use of privacy sensitive information like email addresses makes it difficult to openly test and validate. Our reliance on only publicly obtainable data provides more transparency and utility with respect to a broad range of network measurement scenarios.

3.7 Conclusions

In this chapter, we have proposed a discriminative model of host to address bindings based on preferential attachment at the level of individual addresses. We
have also developed efficient and accurate approximations of this model that permit reliable inferences to be made using real-world data. Our validation focused on the use of anonymised DHCP logs collected at a campus wireless network over the course of four months. We have demonstrated that the number of hosts corresponding to an address is well modelled by the number of times that the address appears. Moreover, this relationship can be expressed as simple mappings between the probability of binding to a new host and the $n^{th}$ appearance by an address.

With respect to the five critical requirements that we developed earlier in Chapters 1 and 2, the major unaddressed limitation at this stage of our proposed methodology is the constraint of only using address characteristics in order to infer host behaviour. In particular, we have yet to establish whether real-world intrusion alerts sufficiently capture enough information about malicious hosts such that their bindings to alerted addresses can be reliably inferred using the model developed in this chapter, while preserving privacy and enabling persistent tracking over long periods of time as well. That is what we specifically investigate in the next chapter of this thesis.
This page intentionally left blank.
Chapter 4
Self-Similar Characteristics of Intrusion Alerts

4.1 Introduction

The goal of this chapter is to investigate the observable characteristics of malicious addresses since the five critical requirements we discussed in earlier chapters constrain us in terms of only using address characteristics to reliably infer host behaviour. As a result, we are interested in the persistent and periodic behavioural patterns of malicious addresses in order to detect any spatial or temporal locality that could be used by the proposed model of host-to-address bindings from the previous chapter. For example, periodic and persistent scanning activity originating from malicious addresses can manifest in long-range dependence, whereby observations during a given measurement interval have strong correlations with those of an earlier interval. Moreover, statistical self-similarity may manifest in scale invariant and skewed distributions of parameters like the aggregate length of appearances by alerted addresses. Our guiding hypothesis is that network intrusion alerts exhibit self-similar characteristics. We present empirical evidence in favour of this hypothesis by analysing a large repository of IDS alerts collected from over 1600 networks distributed throughout the world.

Our analysis focuses on three characteristics that have direct implications for any kind of predictive modelling with regards to dynamic host-to-address bind-
ings. First, we study the persistence of malicious addresses in terms of their aggregate length of appearances, which we refer to as their \textit{lifetime}. We present evidence which shows that the lifetime distribution of malicious addresses detected by a large number of IDSs is heavy-tailed in accordance with a power-law. Furthermore, we find that this relationship does not change when comparing vastly different volumes and categories of intrusion alerts across different networks. Second, we find that the distribution of malicious addresses over the IP address space remains unchanged despite the widely varying lifetime of addresses. In other words, the same regions in the IP address space appear to simultaneously generate short-lived intrusion attempts along with persistent ones. Finally, we present evidence which shows that the number of malicious addresses per unit time follows a scale invariant diurnal cycle (with a corresponding long-range dependence) regardless of the volume or category of intrusion alerts analysed.

The remainder of this chapter is structured as follows. Section 4.2 describes the contributors and properties of the intrusion logs analysed while the methodology of our analysis is explained in Section 4.3. We present the lifetime distribution of malicious addresses in Section 4.4 and their spatial distribution in Section 4.5. Evidence pointing to the scale invariant diurnal cycle is presented in Section 4.6. Finally, we discuss the implications and relevance of our findings with respect to the literature in Section 4.7 before concluding this chapter in Section 4.8.

4.2 Intrusion Alerts for Analysis

The data set we analyse is from DShield, and spans the 14 day period beginning at midnight 1st January 2005 and ending at midnight 15th January 2005. DShield is a large-scale collaborative intrusion detection community, which aims to identify and detect new vulnerabilities as well as generate relevant blacklists
4.2 Intrusion Alerts for Analysis

Table 4.1: Sample rows from the DShield logs (showing only the subset of columns that we have used). Each row (or alert) indicates an intrusion attempt made by the listed Src IP (unaltered) on the targeted Dst Port (unaltered) as observed and classified by the corresponding subscriber’s IDS/firewall (anonymised by the Sub ID attribute). Date and Time values are in GMT format.

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Sub ID</th>
<th>Src IP</th>
<th>Dst Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-01-01</td>
<td>4:26:17</td>
<td>3621</td>
<td>0.0.0.0</td>
<td>135</td>
</tr>
<tr>
<td>2005-01-02</td>
<td>0:49:23</td>
<td>0</td>
<td>34.15.9.8</td>
<td>445</td>
</tr>
<tr>
<td>2005-01-03</td>
<td>9:33:12</td>
<td>65</td>
<td>1.6.16.1</td>
<td>139</td>
</tr>
</tbody>
</table>

There was a total of 1679 DShield subscribers that contributed IDS alerts and firewall logs collected at the gateway of their networks during the two week period that we investigate. DShield provides only minimal common information merged from the various alert streams of its subscribers and uses GMT timestamps. Sample rows from our data set are shown in Table 4.1. We used a subset of the full collection of columns comprising only Date, Time, Subscriber ID, Source IP and Destination Port values, as other attributes were not relevant to our analysis. The Source IP is not anonymised and consequently, the underlying distributions are fully preserved. This is a major advantage of analysing DShield logs due to the high fidelity of source addresses observed from a large number of vantage points in the form of subscribers’ IDSs/firewalls. There were a total of 11,888,272 unique source addresses in this data set containing over 484 million alerts throughout the 14 days.

The contributions of individual DShield subscribers are far from equal in terms of the fraction of total alerts in the data set. Figure 4.1 shows the Cumulative Distribution Function (CDF) after ranking subscribers in ascending order of their
Figure 4.1: Contributions of individual subscribers to the DShield repository in terms of the fraction of total alerts. The knee point is at the 10th percentile mark on the bottom right-hand corner, which covers 1625 “lightweight” subscribers while the remaining 90% of total alerts originate from just 54 “heavyweight” subscribers.

respective contributions. The imbalance is quite striking, as demonstrated by the knee point being at the 10th percentile mark that includes 1625 subscribers out of a total of 1679. This leaves 54 subscribers that we henceforth refer to as “heavyweight”, since they collectively contribute almost 90% of all the alerts in our data set. The remaining 1625 subscribers are correspondingly referred to as “lightweight”.
4.3 Analysis Methodology

We measure the statistical self-similarity of network intrusion attempts along three orthogonal dimensions corresponding to malicious addresses found in the DShield alerts. First, we study the temporal self-similarity of malicious addresses as measured by their lifetime of appearances. Second, we look into the spatial distribution of malicious addresses and whether they can be distinguished by way of scanning categories related to their activities. Third, we investigate the volumetric self-similarity in terms of long-range dependence across multiple measurement intervals. In all three cases, we utilise simple but robust statistical hypothesis tests, which can address the inherent uncertainty and noise that arise in intrusion alerts. We follow the regression testing approach that has been found to be both robust and efficient in related studies of this type [3, 110]. Our analysis proceeds via the following steps, which are subsequently described in detail.

1. Aggregation into Intervals
2. Classification of Addresses
3. Definition of Lifetime
4. Address Distribution
5. Diurnal Cycle

4.3.1 Aggregation into Intervals

The first step of our analysis aggregates alerts into fixed 10-minute intervals and assigns a single time-stamp corresponding to the end of the intervals. The goal is simply to discretise the continuous time-scale in such a way that it facilitates prompt processing while preserving a fine-grained temporal resolution. The choice of 10-minute windows is appropriate in our opinion for two reasons. First, Katti et al. [70] found that 90% of intrusion attempts from malicious addresses arrive within a minute of each other and furthermore, 95% of attempts are within 10
minutes of each other. This finding is applicable to our data set since it is a sub-
set of the one studied by them. As a result, aggregation of alerts into 10-minute
intervals will collect nearly all intrusion attempts from addresses corresponding
to specific attack events. Second, 10 minutes is a suitable resolution for the detec-
tion of hourly and daily cycles that may be present in the logs. This gives us six
data-points for each hour and 144 data-points for each day, which we consider to
be of adequate precision.

4.3.2 Classification of Addresses

Throughout our analysis, we separate the intrusion logs contributed by heavy-
weight subscribers from those contributed by lightweight subscribers. There is a
large difference in the volume of alerts from these two categories of subscribers
as mentioned previously, and the goal of this separation is to determine if the
underlying statistical characteristics change according to the volume of intrusion
attempts reported by subscribers. Another dimension along which we examine
whether the underlying characteristics change or not is the type of intrusions at-
tempted. For this purpose, we classify malicious addresses in each interval into
one of four categories as defined below.

- **Horizontal** - addresses that scanned a common destination port across mul-
tiple subscribers’ networks,
- **Vertical** - addresses that scanned multiple destination ports but only on one
subscriber’s network,
- **Static** - addresses that scanned only one destination port on only one sub-
scriber’s network, and
- **Hybrid** - addresses that scanned multiple destination ports across multiple
subscribers’ networks.
These categories have practical relevance. Horizontal addresses have breadth (i.e., multiple subscribers) but no depth (i.e., just one destination port) to their scanning activity. They are most likely to constitute machines that are actively trying to propagate malware by identifying and exploiting a specific vulnerability. Vertical addresses have depth but no breadth to their scanning activity. These might be malicious addresses that are using a hitlist and targeting one subscriber’s network in order to identify a range of common vulnerabilities. Static addresses most likely constitute zombies instructed to carry out a DDoS attack on one subscriber’s network as they lack both breadth and depth of scanning activity. They perhaps also incorporate a certain number of addresses that are falsely identified (noise) as malicious by various IDSs. Finally, Hybrid addresses have a mixture of breadth and depth, and probably comprise machines that are infected with different kinds of malware (from time to time) such that their scanning activity appears to be mixed.

4.3.3 Definition of Lifetime

We define lifetime as the aggregate number of appearances in discrete time intervals. So, if an address appears in a given interval, we count that as one appearance regardless of the number of alerts it may have generated during that interval. This is consistent with our attempt to discretise the time-scale. Moreover, the result of counting an address’s appearances (at any temporal resolution) is a measure of its persistence, which flattens out the inter-arrival times and assigns an aggregate value for the lifetime of that address [124]. The advantage of measuring persistence in this manner is that it allows us to rank all the addresses in ascending order of their lifetimes and study the resulting “count-frequency” distribution, where we use a log-log plot of the number of addresses (denoted \( f \)) that have a given lifetime count (denoted \( x \)) in order to test the existence of a
power-law relationship (e.g., Zipf distribution \[125\]). In essence, this technique determines the scaling exponent \(\alpha\) assuming a power-law relationship between \(f\) and \(x\) of the form given by Equation 4.1.

\[
f = ax^{-\alpha} \quad \forall x \geq 1, \quad \alpha > 1, \quad a \in (0, \infty)
\]  

(4.1)

The scale invariant property (i.e., self-similarity) of Equation 4.1 is that for any value of the constant \(a\), the function resembles itself given the same scaling exponent \(\alpha\). This becomes clearer when taking the logarithm of both sides of Equation 4.1 as shown in Equation 4.2. Consequently, the gradient of the count-frequency distribution in a log-log plot reflects the scaling exponent \(\alpha\), while the y-axis intercept corresponds to \(\log(a)\). So changing the value of \(a\) simply moves this straight line up or down while preserving its downward slope.

\[
\log(f) = -\alpha \log(x) + \log(a)
\]  

(4.2)

### 4.3.4 Address Distribution

Subsequent to the lifetime analysis of malicious addresses, we investigate their distribution in the IPv4 address space after classifying them into different lifetime categories along with the aforementioned four types of intrusion attempts. The goal here is to determine whether the spatial distribution of addresses changes (i.e., diverges from self-similar characteristics) depending on their lifetimes and types of intrusion attempts. We primarily use CDF plots to analyse the distribution of addresses over the IP address space. We also analyse a property called volatility that was defined in our earlier work \[74\]. It measures the cardinality of the set of all addresses observed up to and including a given interval. In essence, volatility depicts the rate at which brand-new addresses appear as a function of time. So, high volatility corresponds to a constant appearance of new addresses.
in each interval that have hitherto not been observed and leads to a lack of spatial locality. Conversely, low volatility means the recurring appearance of virtually the same set of addresses in each interval and leads to a persistent spatial locality.

4.3.5 Diurnal Cycle

The final characteristic we analyse is the existence of periodic patterns and long-range dependence in the number of unique malicious addresses observed during each interval. We use the autocorrelation function to reveal the length of any cycle and the extent of any long-range dependence that may exist. The intuition here is that treating the number of addresses observed in each interval as a time-series and plotting its autocorrelation function at different time-lags will allow us to detect any cyclical patterns. We centre a time-series under investigation about its mean in order to remove any constant bias (i.e., de-trending) before computing the autocorrelation function.

The implication of a diurnal cycle in a de-trended time-series is that it manifests in an autocorrelation function that oscillates with a period equal to 24 hours and a slowly decaying envelope that eventually suppresses these oscillations to zero. The existence of such oscillations usually indicates statistical self-similarity in a volumetric sense [118]. That is, the number of unique malicious addresses for any given interval has a strong likelihood of appearing at a later interval. In a practical sense, such a phenomenon would be a direct result of the activities of malicious addresses from a group of adjacent time-zones, which in turn would imply a regional cluster of malicious hosts.
4.4 Lifetime Distributions

We begin by presenting the count-frequency distribution of the full set of malicious addresses that were reported by all the subscribers in Figure 4.2. As evident, the blue coloured crosses in this log-log plot visually follow a downward linear trend. The linear regression estimate of $\alpha$ is -2.42 with $r^2 = 0.97$ at the 0.01 level of significance. Therefore, we cannot reject the existence of a heavy-tailed distribution for the lifetime of malicious addresses.

The question we now ask is does the heavy-tailed lifetime distribution keep
emerging for vastly different volumes of intrusion attempts experienced? To answer, we present the count-frequency distribution for malicious addresses reported by heavyweight and lightweight subscribers in Figure 4.3. As can be seen, the heavy-tailed lifetime distribution is nearly identical for both groups. Lightweight subscribers lead to an estimated $\alpha = -2.33$ while heavyweight subscribers lead to an estimated $\alpha = -2.41$ with both groups having $r^2 > 0.95$. This is a striking exhibition of self-similarity given the heavyweight subscribers contribute 90% of the total alerts in our data set and the underlying lifetime distribution is unchanged for both groups of subscribers.

Given that the structure of aggregated IP addresses on the Internet has been observed to follow a heavy-tailed distribution for different lengths of the common prefix [110], we now investigate if that result is reflected by the sources reported in the DShield logs. In Figure 4.4 we show the count-frequency distribution of /24 subnets extracted from the set of source IP addresses reported by heavyweight and lightweight subscribers separately. In both cases it is evident that the power-law relationship holds given the closeness of $\alpha$ values to that of Figure 4.2 and the corresponding values of $r^2$.

Finally, we investigate whether the heavy-tailed lifetime distribution changes depending on the type of intrusions attempted. Figure 4.5 presents the count-frequency distribution of Horizontal addresses reported by heavyweight and lightweight subscribers separately while Figure 4.6 does the same for Static addresses (plots for Vertical and Hybrid addresses exhibit the same result). In all cases, we can see clearly that the heavy-tailed distribution exists despite obvious differences in the type of intrusions attempted. Recall that Horizontal addresses are scanning across multiple subscribers while Static addresses are scanning only one subscriber and yet, the lifetime distributions of both types of addresses obey the power-law relationship. Notice that all relationships are significant at the 0.01 level.
Figure 4.3: Count-frequency distribution of the subscriber-specific sets of malicious addresses when ranked in ascending order of aggregate lifetime.
Figure 4.4: Count-frequency distribution of the subscriber-specific sets of /24 subnets in the data set when ranked in ascending order of aggregate lifetime.
Figure 4.5: Count-frequency distribution of the subscriber-specific sets of Horizontal addresses when ranked in ascending order of aggregate lifetime.

(a) Heavyweight subscribers

(b) Lightweight subscribers
4.4 Lifetime Distributions

Figure 4.6: Count-frequency distribution of the subscriber-specific sets of Static addresses when ranked in ascending order of aggregate lifetime.
Figure 4.7: Count-frequency distribution of the subscriber-specific sets of Vertical addresses when ranked in ascending order of aggregate lifetime.
Figure 4.8: Count-frequency distribution of the subscriber-specific sets of Hybrid addresses when ranked in ascending order of aggregate lifetime.
4.4.1 Implication

The practical implication of these findings is that an overwhelming majority of malicious addresses never actually re-appear, despite being classified as suspicious by DShield subscribers. On the other hand, there is a small group of addresses that are highly persistent as observed by DShield subscribers. So this implies that any potential model for predicting host bindings must take into account the lifetime of malicious addresses when extrapolating their past behaviours, and that the prediction window into the future will need to be a function of the lifetime distribution of the addresses observed so far in the past (i.e., the training period). In other words, if all we have seen in the past are short-lived intrusion attempts then our prediction window can at best be the longest lifetime observed and consequently, any resulting blacklists have a low likelihood of being valid beyond that point.

4.5 Address Distributions

In this section we investigate the question - where do malicious addresses tend to come from and how does their distribution over the IP address space vary based on different categories of lifetimes and intrusion attempts? Figure 4.9 shows the CDF of IP address distributions observed by heavyweight and lightweight subscribers separately. In each case, we plot the overall distribution of IP addresses observed by the corresponding group of subscribers along with the distribution of addresses classified into three different categories of lifetimes. Namely, these categories are single-attempts (appears in one interval), short-lived (appears in between 2 and 100 intervals) and long-lived (appears in more than 100 intervals). As evident in Figure 4.9, the distributions appear to overlap considerably. In the case of heavyweight subscribers, the different distributions are almost indistinguishable while the overall shape of the CDF plots are identical for both heavy-
weight and lightweight subscribers. So this implies that regardless of the lifetime of malicious addresses, they appear from the same regions of the IP address space.

Switching focus to whether these distributions change when viewed by different categories of intrusion attempts, we plot the results in Figure 4.10. Once again it appears that the IP address distributions remain unchanged across the heavyweight/lightweight subscriber split and across different types of intrusion attempts. In fact, the visual overlap of distributions is more pronounced for both heavyweight and lightweight subscribers in this case than in the case of different lifetime categories previously.

Finally, we investigate the volatility of IP addresses observed in a cumulative manner for different categories of intrusion attempts. The results are plotted in Figure 4.11. The linear curves for all of the different cases indicate that the number of brand-new addresses observed (that have not been seen so far) in each interval is virtually constant and remains unchanged across the categories of intrusion attempts. In other words, there is a steady level of high volatility in the set of IP addresses observed thus far as we cumulatively grow that set over each interval until reaching the end of the data set (the normalisation factor for the CDF plots).

**4.5.1 Implication**

The practical implication of these findings is that there is no notion of spatial locality for IP addresses when analysing what may appear to be bursts of intrusion attempts. This is because single-attempts, short-lived and long-lived malicious addresses all originate from the same regions of the IP address space such that we cannot persistently distinguish address aggregates that are malicious from those that are not. In addition, the distribution of addresses are identical when
Figure 4.9: Distribution of malicious addresses over the IPv4 space for different categories of lifetimes.
4.5 Address Distributions

Figure 4.10: Distribution of malicious addresses over the IPv4 space for each category of intrusion attempts.
Figure 4.11: Normalised rates of new address appearances for each category of intrusion attempts and the overall unclassified attempts.
comparing different categories of intrusion attempts against different categories of lifetimes. As a result, this compounds the implication that there is no bounded size of a burst of intrusion attempts (both temporally and spatially) that we could take advantage of when applying our model of host-to-address bindings.

4.6 Diurnal Cycles

The focus of our analysis in this section is the number of unique sources appearing in each interval and whether this exhibits long-range dependence along with cyclical patterns. In Figure 4.12 we present the time-series of the number of malicious addresses reported by all subscribers and the autocorrelation function of the de-trended time-series. In the first instance, there is a visibly detectable diurnal cycle where the number of addresses appearing in each interval peaks periodically above a constant baseline. The autocorrelation function confirms that the period of this pattern is 24 hours as evident from its peaks occurring at time-lags of 24 hours and its integer multiples. We can also observe a long-range dependence in the fact that the autocorrelation function decays slowly towards zero at time-lags of more than a week. Therefore, this establishes the existence of a diurnal cycle with long-range dependence in the number of malicious addresses appearing per interval.

We now examine whether the diurnal cycle and corresponding long-range dependence change for the different groups of subscribers with vastly different volumes of intrusion alerts experienced. The time-series of the number of addresses appearing in each interval corresponding to heavyweight and lightweight subscribers are plotted in Figure 4.13 along with the autocorrelation functions corresponding to the de-trended time-series. Once again, we find that heavyweight and lightweight subscribers exhibit similar characteristics. In this case, the time-series corresponding to lightweight subscribers is significantly smoother than
Figure 4.12: Time-series and corresponding autocorrelation function of the number of unique addresses per interval reported by ALL subscribers in the data set.
that of heavyweight subscribers and that is reflected in their respective autocorrelation functions. Nevertheless, both of them display a clear diurnal cycle with long-range dependence.

As in the earlier section, we investigate whether the aggregation of IP addresses impact this diurnal cycle or not. In Figure 4.14 we plot the time-series corresponding to the /24 subnets extracted from the set of unique sources appearing in each interval along with the autocorrelation function. The same diurnal cycle with long-range dependence still exists albeit at a smaller scale due to the lower volume of /24 aggregated addresses compared to the full /32 addresses.

Finally, we plot the time-series and autocorrelation functions corresponding to Horizontal and Static addresses in Figures 4.15 and 4.16 respectively (plots for Vertical and Hybrid addresses exhibit the same trend also). In all cases, we find that the diurnal cycle is present with a long-range dependence in the autocorrelation function. The striking self-similarity is evident by the considerable difference in volume of addresses reported under each category of intrusion attempts. Static addresses are far more voluminous than horizontal addresses in any given interval but both of them still exhibit the same diurnal cycle with long-range dependence.

### 4.6.1 Implication

The implication of an omnipresent diurnal cycle in the number of unique addresses appearing per interval is that it correlates with human activities from a handful of adjacent time-zones. This is unexpected given that DShield subscribers are distributed around the world and spread across nearly all the different time-zones. This is further compounded by the long-range dependence of over a week, which means that any model for predicting future attacks must account for a considerable amount of history. Such long-term tracking require-
Figure 4.13: Time-series and corresponding autocorrelation functions of the number of unique malicious addresses per interval reported by Heavyweight and Lightweight subscribers.
Figure 4.14: Time-series and autocorrelation functions of the number of unique /24 subnets per interval reported by Heavyweight and Lightweight subscribers.
Figure 4.15: Time-series and autocorrelation functions of the number of unique Horizontal addresses per interval reported by Heavyweight and Lightweight subscribers.
Figure 4.16: Time-series and autocorrelation functions of the number of unique Static addresses per interval reported by Heavyweight and Lightweight subscribers. Notice how Static addresses are far more voluminous compared to Horizontal addresses and yet the same diurnal cycle is present in both categories.
Figure 4.17: Time-series and autocorrelation functions of the number of unique Vertical addresses per interval reported by Heavyweight and Lightweight subscribers.
Figure 4.18: Time-series and autocorrelation functions of the number of unique Hybrid addresses per interval reported by Heavyweight and Lightweight subscribers.
ments generate obvious challenges for efficient storage and processing of logs in order to utilise long cycles for more accurate forecasting of future intrusion attempts. It is also possible that cycles with periods shorter than 24 hours exist and are superimposed with the dominant diurnal cycle to create a signal with a much broader spectrum than the one we have uncovered here. Nevertheless, any recurring pattern will lead to this kind of long-range dependence.

4.7 Related Work

The self-similarity idea for our work stems from the seminal paper by Leland et al. [113], which rigorously established the fractal characteristics of Ethernet traffic. There has been considerable work done with regards to self-similarity in various aspects of networking since their groundbreaking research. The most relevant of these is the work of Kohler et al. [110] that found self-similar properties in the structure of aggregated IP addresses observed at various points on the Internet. Their work was extended to modelling the structure of malicious sources on the Internet by Barford et al. [111]. Despite dealing with statistical self-similarity, neither of them investigated the three specific characteristics that we have studied in this chapter.

There have been two notable longitudinal studies carried out in the past that established certain characteristics of scanning activity on the Internet by analysing DShield logs. The earliest study in [20] analysed four months of DShield logs from 2001-2 and found that a small collection of addresses were responsible for the majority of scanning activity in any given month. They also found that various well known worms at the time had persisted long after their initial release, which showed signs of coordinated behaviour among the set of globally alerted addresses. Their findings were extended by Chen et al. in [101], who found that 80% of the malicious addresses occupy only 20% of the IP address space after
analysing DShield logs for 402 days spanning from 10th November 2004 till 10th September 2006. They also found that a majority of the addresses only appeared on one day out of the possible 402 days in their data set. Both of these studies present vague evidence pointing towards self-similar characteristics of network intrusion attempts without rigorously examining this hypothesis in the manner that we have done. Moreover, our novel focus has been on finding direct implications of self-similar characteristics on the predictability of host bindings.

Two other studies that analysed DShield logs appear in [70] and [63]. Both were interested in characterising the scanning activity of malicious addresses for more efficient [70] and relevant [63] blacklist generation. Katti et al. in [70] found that 40% of the scanning activity is correlated in terms of one alerted address targeting multiple subscribers while 20% of the malicious addresses engaged in such correlated activities. They also found correlated groups of around 6 subscribers with a corresponding set of common attackers that changed very little over time. This formed the basis of their argument for sharing IDS alerts with only the fellow subscribers in a correlated group rather than the whole set of 1700 subscribers. Their argument was extended by Zhang et al. in [63], who used a modified version of Google’s PageRank algorithm to determine how relevant an alerted address was to a subscriber based on the past behaviours of both. The outcome was a blacklisting service customised for each subscriber to achieve improved hit rates. Our findings have implications for both of these studies and their respective strategies for predicting future attacks. The lack of spatial locality in the IP address space regarding different categories and lifetimes of intrusion attempts means a temporal upper bound must be placed on the validity of any blacklist, which was not taken into account by these two studies.

Finally, a detailed study of scanning traffic by Pang et al. appears in [19]. Based on passive and active packet capture from scan related flows, they found that the overwhelming majority of destination ports scanned were well known
vulnerabilities like port 80 (HTTP), port 135 (RPC) and port 139 (NetBIOS). They also found a diurnal cycle in peak scanning traffic that implies a causal relationship with human activity and parallels our findings of an omnipresent diurnal cycle with long-range dependence. In comparison to [19], our definition of the four categories of intrusion attempts is deliberately oblivious to specific vulnerabilities, since we are interested in understanding collective trends and not specific types of scanning exploits that are most prevalent on the Internet.

4.8 Conclusions

In this chapter, we examined the hypothesis that network intrusion alerts exhibit self-similar characteristics. We analysed logs from the DShield repository of intrusion alerts corresponding to the first two weeks of January 2005. Our findings support this hypothesis by primarily establishing the fact that hosts corresponding to network intrusion attempts display identical statistical characteristics regardless of the volumetric, spatial and categorical resolution of observations. For example, the lifetime distribution of malicious addresses (which is defined as their aggregate length of appearances) is heavy-tailed and follows a power-law. We found that this remains the same for different groups of subscribers who contribute vastly different volumes of alerts to the DShield repository. Our second finding that supports this hypothesis is the broadly identical distribution of malicious addresses in the IP address space regardless of different categories of lifetimes and intrusion alerts. The third and final piece of evidence that leads us to conclude in favour of the hypothesis is the existence of a diurnal cycle with long-range dependence in the number of unique addresses appearing per unit time, which also does not change despite varying the scale in which we observe this cycle.

The collective implications of our findings are significant when it comes to
4.8 Conclusions

predicting future activities of malicious addresses based on their past intrusion attempts and thus, relying on blacklisting as a baseline countermeasure. The existence of a heavy-tailed distribution in the lifetime of addresses means that the overwhelming majority of sources that are blacklisted (or at least alerted) by IDSs never actually re-appear. The number of highly persistent sources is tiny in comparison. This issue is compounded by the broadly identical distribution of all sources in the IP address space regardless of their lifetimes and the types of intrusion attempts. Therefore, the notion of spatial locality does not hold when it comes to temporally associating bursts of malicious activity with IP address aggregates and any corresponding blacklisting of subnets becomes flawed. The scale invariant diurnal cycle along with its long-range dependence over many days implies strong correlation with human activities in a handful of adjacent time-zones (e.g., continental Europe). Given that our data set comprises networks distributed throughout the world and across nearly all the different time-zones, it is surprising to find such an omnipresent cyclical correlation. Nevertheless, any kind of dynamic blacklisting must take into account such a recurring time-of-day effect.

At this stage of our methodology for estimating the number of malicious hosts corresponding to an alerted address, we have established that intrusion alerts capture necessary address characteristics to detect malicious host behaviour (e.g., different kinds of port-scanning), but are not sufficient for reliable identification and persistent tracking of hosts due to the absence of spatial and temporal locality. Therefore, the question that remains unanswered is whether there is any other form of readily available measurements that can help reduce these uncertainties or not? In particular, can we filter unreliable or misleading addresses that appear in alerts as a result of devious behaviour (e.g., spoofing)?

In the next chapter, we propose the use of active probing to complement the passive approach to identification and tracking that has been used in this chapter,
in order to improve the reliability of malicious host identification based only on
the address characteristics. It is also worth noting here that apart from the techni-
cal requirements of reliable identification and persistent tracking, the remaining
critical requirements of privacy preservation, reproducible validation and broad
applicability are well satisfied by intrusion alerts. Hence the justifiable need to
overcome their deficiencies by correlating with other forms of independent mea-
urements as opposed to discarding the use of intrusion alerts entirely.
Chapter 5
Correlation of Intrusion Alerts with Active Measurements

5.1 Introduction

The probabilistic model of host-to-address bindings that we have developed in Chapter 3 requires that observed addresses be characterised in terms of their discrete appearances. However, our subsequent findings from Chapter 4 indicate that intrusion alerts alone are not sufficient for use with this model due to the inherent uncertainty arising from the loss of spatial and temporal locality, spoofed sources and volatility of appearances that collectively make it unreliable to infer the number of hosts bound to an alerted address. Given the passive nature of intrusion alerts, the key question that needs to be addressed is: can active probing at the network layer improve the reliability and persistence of host identification under the constraints described in the first two chapters of this thesis?

The goal of this chapter is to answer this question by proposing a novel methodology for analysing independently gathered passive and active measurements such that our model of host-to-address bindings can be used to draw robust inferences about the number of hosts bound to individual malicious addresses. Our analysis is carried out in two parts. First, we examine the compatible statistical characteristics of the two independently collected data sets, namely intrusion alerts from DShield.org [71] and ping responses gathered by the University of
Southern California [126]. Second, we apply the model of host-to-address bindings in order to infer the distribution of malicious hosts on a per address basis across the globally visible IPv4 space. In so doing, we found that more than 80% of the alerted addresses bound to multiple hosts and that their distribution is highly non-uniform across the global IPv4 space.

The remainder of this chapter is structured as follows. We present the details of the two data sets in Section 5.2. Our approach to analysing these data sets is presented in Section 5.3. Then we examine their statistical characteristics and compatibility in Section 5.4. The subsequent results of applying the model of host-to-address bindings are presented in Section 5.5. Finally, we discuss the implications of our findings in Section 5.6 along with a discussion of other relevant results from the literature before concluding the chapter in Section 5.7.

5.2 Data Sets Analysed

The two data sets that we have analysed both correspond to the fortnight starting on midnight 18 August 2010 and ending on midnight 2 September 2010. The DShield data set has an identical format as that of the one presented earlier in Table 4.1. The ping response data set was considerably larger (around 70 GB compressed) due to the repeated probing of randomly selected global addresses every 11 minutes. These independent targets constituted 1% of the allocated IPv4 global pool. The detailed methodology of this probing can be found in [59], while various meta-data corresponding to the exact data set that we analysed can be found in [126] and [127]. Figure 5.1 summarises the data sets and our overall approach to analysing them.
5.2 Data Sets Analysed

Globally distributed networks
8+ million alerted addresses
1000+ separate IDSs
~30 GB zipped archive

DShield Alerts
ICMP Echo Replies

DShield Alerts
ICMP Echo Replies

2010-Aug-18_12am until 2010-Sep-02_12am
28,516 common addresses that were real, malicious and active

Distribution of malicious hosts

Estimated # of Hosts
IPv4 Address Space

Probed 1% of global IPv4 pool
2+ million responsive addresses
11 minute ping intervals
~70 GB zipped archive

Figure 5.1: Outline of our approach to estimating the number of malicious hosts by correlating intrusion alerts and ping responses. The top part entails independent collection and analysis of the respective data sets. The middle part entails correlation of these data sets in order to retrieve their intersection. The bottom part combines the upper two by applying the proposed model of host-to-address bindings for estimating the number of hosts corresponding to each globally visible malicious address.
5.3 Our Analysis Approach

The problem we are specifically addressing in this chapter is that of reliably identifying host behaviour based only on address characteristics. To that end, the data sets that we have chosen provide a number of advantages in our analytical approach. First, actively pinging globally visible addresses filters out spoofed, unbound, unallocated and inactive addresses in an efficient and informative manner. Second, a globally distributed repository of intrusion alerts provide multiple vantage points for monitoring, which help to temporarily identify malicious addresses. The combination of these two advantages leads to a scenario whereby we are able to reliably identify malicious addresses over a long period of time. Finally, the most important advantage is that the intersection of these two data sets is directly applicable to our model of host-to-address bindings such that we can infer the number of hosts per address while satisfying the five critical constraints of privacy preservation, reliable identification, persistent tracking, broad applicability and reproducible validation.

5.3.1 Processing Ping Responses

As described in Section 5.2, we have used a publicly available Internet Control Message Protocol (ICMP) ping probe data set, in which the probes were issued on an 11 minute interval. Consequently, this 11 minute timing period became the basis of our subsequent analysis. Responses from a probed address during each 11 minute interval can be one of three types in general, as abstracted from the ICMP specification [128]. First, an address can respond positively within the 11 minute probe interval, which indicates that it is allocated, reachable and actively bound to a host during that interval. We interpret such a probe interval containing a positive response from an address as part of an ON period. Second, an address may not respond at all during the 11 minute probe interval (i.e., timeout
Table 5.1: Chronologically ordered sequence of records from the ping response data set corresponding to a probed IPv4 address that demonstrates the three types of responses possible.

<table>
<thead>
<tr>
<th>Response</th>
<th>UTC Timestamp</th>
<th>Probe Address</th>
<th>Reply Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>2010-08-18 01:06:15</td>
<td>98.112.82.64</td>
<td>98.12.82.64</td>
</tr>
<tr>
<td>Positive</td>
<td>2010-08-18 01:17:24</td>
<td>98.112.82.64</td>
<td>98.12.82.64</td>
</tr>
<tr>
<td>Negative</td>
<td>2010-08-18 01:28:36</td>
<td>98.112.82.64</td>
<td>69.50.49.40</td>
</tr>
<tr>
<td>No-reply</td>
<td>2010-08-18 01:39:44</td>
<td>98.112.82.64</td>
<td>NULL</td>
</tr>
<tr>
<td>Positive</td>
<td>2010-08-18 01:50:51</td>
<td>98.112.82.64</td>
<td>98.12.82.64</td>
</tr>
</tbody>
</table>

due to no reply). This can indicate one of several possibilities: namely that the address is unallocated, inactive, unbound, or unreachable. Differentiating among these possibilities is not necessary in our analysis since we combine all of them into an OFF period corresponding to the address. Finally, a probed address can be indicated to be unreachable or unallocated by a different address within the 11 minute probe interval. This is a negative reply that we also interpret as part of an OFF period for the corresponding probed address. These three types of replies are detailed in Table 5.1. Also note that we utilised the 1-loss repair technique [59] in order to be robust against network outages that potentially caused negative or no replies.

5.3.2 Intersection of Ping Responses and Intrusion Alerts

The 11 minute probe interval for the ping response data set provided an adequate resolution for discretising the time-scale of our two week measurement window (i.e., there were 1,833 probe intervals during the fortnight under observation). Prior measurements in the literature indicate that an 11 minute interval is sufficient to observe typical volatile address occupancies of 75 minutes [43] and 81 minutes [59]. This allowed us to aggregate all intrusion alerts falling into any given 11 minute probe interval corresponding to an address as an ON period of malicious activity. Similarly, the remaining probe intervals were considered as
OFF periods. Our method of intersecting these time-series is illustrated in Figure 5.2.

All addresses were independent due to their random selection for probing, which meant that the individual time-series corresponding to a set of addresses were amenable to parallel processing in order to improve running times. This is important when considering the fact that there were over two million probed addresses and over eight million alerted addresses that resulted in an aggregate data set of more than 100 GB compressed. We found that there were 28,516 common addresses that appeared in both data sets. This is indeed a large sample of addresses that were detected via two different measurement methods and two independent groups of observers. Therefore, it provides a substantial body of evidence that the findings reported in the rest of this chapter, based on this common set of addresses, are representative of the broader Internet and practically actionable in terms of better countermeasures.

5.4 Statistical Characteristics

It is necessary to establish whether the two data sets have compatible statistical characteristics prior to applying the model of host-to-address bindings such that any subsequent inferences are known to be valid and reliable. In that regard, we now investigate the self-similar properties exhibited by the common set of 28,516 addresses found in both data sets. Our analysis is rooted in the following hypothesis.

**Hypothesis 5.1.** *A hidden host causing intrusion alerts to be generated from malicious activity is also the same host that is concurrently responding to the ICMP echo probes.*

In other words, our objective is to find out whether these two data sets are in fact different aspects of the same underlying process or not. The motivation for
Figure 5.2: Illustration of steps involved in computing the intersection of our data sets. (a) We begin with an alternating sequence of ON and OFF intervals from the ping responses. (b) Then the intrusion alerts are superimposed to their corresponding probe intervals. (c) Finally, the intersection is obtained for those probe intervals where both an intrusion alert was raised and a positive ping response was received.
doing so is that uncertainty arises from the fact that we can only observe address characteristics instead of host behaviour. Our approach to testing this hypothesis comprises three ideas as described below.

- **State Transitions**: each address undergoes distinct OFF-to-ON and ON-to-OFF transitions as per the state machine representation of its time-series illustrated in Figure 3.1. These transitions provide the means by which to directly count or measure the number of appearances made by an address. Hence, an important test of Hypothesis 5.1 involves checking whether the two independent data sets provide statistically indistinguishable state transitions among the common set of addresses or not.

- **Sojourn Intervals**: the sojourn time of an address in each state is assumed to be an i.i.d sample. This assumption needs to be validated while also checking that the two different data sets do actually produce statistically dissimilar sampling distributions in order to falsify Hypothesis 5.1. Moreover, either the ON period or the OFF period distribution is required to be a power-law if the underlying assumption that hosts bind to addresses via preferential attachment is to remain valid [118]. The general form of the power-law we are testing is the following:

\[
\log(f) = -\alpha \log(x) + \log(a)
\]  

where \(f\) is a count-frequency distribution for the ranked categories in \(x\) and \(\alpha\) is the invariant scaling exponent that parameterises the distribution.

- **Long-range Dependence**: in the case of statistical self-similar characteristics, it is necessary to establish the presence of long-range dependence among the different time-series corresponding to the different data sets. We are particularly interested in identifying any cyclical patterns by way of comparing the autocorrelation function of each time-series in order to
Figure 5.3: Frequency distribution of the OFF-to-ON transitions in each data set. The slope of the linear regression is significant at the 0.01 level in each case.

test whether it decays slowly over a long period of time or not.

Figures 5.3 and 5.4 present the frequency distributions of the two state transitions. It is evident that the distribution in each data set obeys a power-law from Equation 5.1 with the scaling exponent $\alpha$ having a value of around -2.3 across all four cases (significant at the 0.01 level assuming log-normal error distribution). Moreover, the 99% confidence interval estimates for $\alpha$ from Figures 5.3 and 5.4 overlap sufficiently such that we cannot reject the hypothesis that the two independent observations correspond to the same underlying process. Also note that the coefficient of determination $r^2$ explains more than 90% of the variance in all cases that exist between the observed data and the corresponding slope of linear
Figure 5.4: Frequency distribution of the ON-to-OFF transitions in each data set. The slope of the linear regression is significant at the 0.01 level in each case.
5.4 Statistical Characteristics

Figure 5.5 illustrates the existence of a power-law in the OFF period distribution across both data sets. The diverging distributions for ON periods are plotted in Figure 5.6. The 99% confidence interval estimates for the scaling exponent \( \alpha \) for the OFF periods overlap sufficiently such that we cannot reject the hypothesis that the two OFF period distributions arise from the same underlying process.

It is necessary to investigate the reason for the divergent distributions of ON periods. A closer look at the unique number of active addresses in both data sets concurrently in any given probe interval sheds light on a possible cause. As evi-
Figure 5.6: Frequency distribution of ON periods in each data set. The slope of the linear regression is significant at the 0.01 level in each case. Also note that the length of each ON period is measured in 11 minute probe intervals.
dent from the time-series and corresponding autocorrelation in Figure 5.7, there are far more active addresses that reply positively to an ICMP echo probe in any given probe interval than there are malicious addresses that generate intrusion alerts during the same interval. As a result, this must necessarily affect the distribution of ON periods corresponding to the two data sets. However, note the existence of an identical 24 hour diurnal cycle in both data sets, as revealed by the oscillating and slow-decaying autocorrelation function. This provides further confidence that the correlation of these two data sets enables us to reliably identify and track the underlying malicious host(s) corresponding to each independently observed address.

5.5 Distribution of Malicious Hosts

We are now in a position to apply the model of host-to-address bindings to the intersection of the two data sets in order to infer the distribution of malicious hosts across the IPv4 address space. This requires computing the Binomial distribution of Equation 3.2 by utilising the two different mapping functions from Equations 3.3 and 3.5. As per the validation of these mapping functions in Chapter 3, we can substitute the following equations with the estimated $\alpha$ and $\lambda$ values in place of Equations 3.3 and 3.5 respectively. Note the switch to base 10 from natural base $e$ for numerical convenience.

\[ p_n = 10^{-0.11} \cdot n^{-0.59} \]  
\[ p_n = 10^{-1.0172} \cdot 10^{-0.0019} n \]

The observable independent variable in our model is the number of appearances by an address, denoted $n$. The frequency distribution of this measurement found in the intersecting data set is plotted in Figure 5.8. The total number of
Figure 5.7: Time-series and corresponding autocorrelation function of the number of active addresses per 11 minute probe interval in each data set. Note that the period uncovered by the autocorrelation function corresponds to approximately 24 hours (as measured in the number of 11 minute probe intervals).
addresses is 28,516 (as mentioned earlier). However, recall that each address is independent and thus, the frequency distribution of the appearance count of addresses is in fact a distribution of the sample sizes that we are using to draw our inference. As evident from Figure 5.8, there are a large number of samples in the region $10^0 \leq n \leq 10^2$. In fact, more than 95% of the total addresses reside in this region. This provides sufficient statistical power such that we can be confident that any corresponding effect size will be a significant finding that is not due to pure chance alone.

In Figures 5.9 and 5.10 we plot the results of applying the exponential variant
of the model of host-to-address bindings (i.e., Equation 5.3). We observe that more than 80% of the addresses bind to multiple hosts as shown on the histogram of Figure 5.9, while the distribution of the number of hosts is highly non-uniform across the IPv4 space, as shown in Figure 5.10. The results are similar but more volatile when applying the power-law variant of the model (i.e., Equation 5.2), as illustrated in Figures 5.11 and 5.12. In fact, the power-law variant leads to an inference that more than 90% of the addresses bound to multiple hosts with a similar non-uniform distribution across the global IPv4 space as the exponential variant. The two histograms collectively provide an estimate that the expected number of hosts per address is between two and six, since the majority of the addresses fall in that interval.

5.6 Implications and Related Findings

Given the aforementioned results of applying our model to an intersected data set comprising intrusion alerts and ping responses, we now consider their practical implications. First of all, it is clear that there are hidden details buried deep in disparate data sets that only become visible once they are correlated. In purely statistical terms, it is an important validation of the expectation that randomly selecting and pinging globally visible addresses will garner responses from hosts that engage in malicious activity, and consequently show up in an independent repository of globally distributed IDS/firewall logs collected from more than a thousand network gateways. This implies that future experiments on actively probing the Internet can more specifically target addresses that have had a known malicious history based on archived intrusion alerts.

Then there are implications of address aliasing for various counter-measures that rely on a persistent static binding between addresses and hosts. In particular, the practice of static blacklisting is prone to being ineffective when the blocked
Figure 5.9: Histogram of the estimated number of hosts per address in the intersected data set based on the exponential variant of our host-to-address binding model.
Figure 5.10: Scattered distribution of the number of hosts per address (exponential variant of our model) across the IPv4 space with respect to the uniform baseline of static blacklisting. The gaps in the baseline indicate address prefixes that were either unallocated or reserved for special use [129].
Figure 5.11: Histogram of the estimated number of hosts per address in the intersected data set based on the power-law variant of our host-to-address binding model.
Figure 5.12: Scattered distribution of the number of hosts per address (power-law variant of our model) across the IPv4 space with respect to the uniform baseline of static blacklisting. The gaps in the baseline indicate address prefixes that were either unallocated or reserved for special use [129].
addresses tend to be highly dynamic and malicious simultaneously [15]. In addition, the blacklisting of long prefixes (e.g., /24 blacklists published by DShield [63]), can lead to collateral damage whenever the usage characteristics of those prefixes accommodate both malicious and legitimate hosts [16]. Our finding provides tangible evidence in support of this by demonstrating that malicious addresses that bind to multiple hosts are not uniformly concentrated across the IPv4 space. As a result, different /24 prefixes are likely to have different levels of infection even though blacklisting them all may generate a uniform level of intrusion alerts. This means that there is significant scope for future blacklisting mechanisms to dynamically correlate data from disparate sources in order to reduce the inherent uncertainties and resolve various ambiguities.

There is related work in the literature that has utilised a similar tactic to us. For instance, Ding et al. [32] recently proposed the correlation of IP flow measurements with DShield intrusion alerts in order to uncover tight knit communities of addresses that form a baseline of activity against which “anti-social” or promiscuous addresses stand out, since they appear to be engaged in numerous communities. However, the dynamic nature of address allocations and translations mean that their approach is not robust against the aliasing and ambiguities that we have highlighted in our findings. Perhaps a hybrid approach is possible whereby our method of inferring the likely number of hosts for a given address can be used to determine whether an address is indeed promiscuous or in fact recycled across multiple hosts.

Furthermore, Wilcox et al. [16] correlated ping responses with spam blacklists while Sinha et al. [15] looked at aggregating different spam blacklists in a dynamic manner. The major finding in both of these studies was that static blacklisting becomes ineffective quickly due to various factors like dynamic addressing, fast flux service networks and rapidly evolving malware life-cycles. The proposed remedies, however, still require a method that resolves address alias-
Correlation of Intrusion Alerts with Active Measurements

ing. Consequently, our methodology of correlating intrusion alerts with ping responses can be used as a baseline signal to improve the accuracy and confidence of various dynamic blacklisting policies (refer to [64] for an extensive survey).

Finally, it is worth considering the implications of privacy and the consequent limitations that are placed on accuracy. The act of correlating two independent sources of data can violate privacy in theory. However, as long as users are not obviously implicated then privacy preservation is adequately provided by ensuring data anonymisation and sanitisation, which have been carried out during our analysis. Although the distribution of malicious hosts across the IPv4 space appears to theoretically violate privacy in the sense that it allows geolocating a malicious host, and thus its corresponding user(s), we do not believe that this can be done with sufficient accuracy to violate privacy in practice. For instance, Triukose et al. [67] have demonstrated the large spatial errors present in such inferences. Moreover, the finding that we highlight is the non-uniform nature of the distribution, which has implications for static blacklisting of fixed length prefixes such as /24. We have not attempted any inference that implicated the latent set of users associated with the observed addresses in any way. Therefore, the privacy preservation requirement that we set out to satisfy has been successfully met by the methodology detailed in this chapter.

5.7 Conclusions

In this chapter, we have applied the probabilistic model of host-to-address bindings to an intersection of intrusion alerts and ping responses. This allowed us to identify thousands of addresses that were active and malicious. We found that more than 80% of these addresses bound to multiple hosts and the distribution of hosts across the IPv4 space was highly non-uniform. This has implications for the ineffectiveness of static blacklisting and the need to pursue more dynamic forms
of countermeasures that can reliably track malicious hosts over many weeks and months. Our work demonstrates that challenging practical scenarios in the context of network management, security and measurement can be handled while still satisfying the legal requirement of privacy protection.

Furthermore, we have addressed the challenge of reliably and persistently identifying malicious host behaviour based only on address characteristics via the correlation of independently gathered passive and active measurements. The collective contribution of this chapter along with the two preceding chapters is that the critical requirements of privacy preservation, reliable identification, persistent tracking, broad applicability and reproducible validation have been satisfied by our proposed methodology for inferring the number of malicious hosts corresponding to an alerted address.

The inferences we have drawn are robust given that they are based on addresses that were allocated, bound, active and malicious simultaneously in two orthogonal forms of measurement carried out independently. The data sets we have used are openly accessible and our model of host-to-address bindings is validated using a public ground truth. The end result is that we have developed a novel methodology for inferring the distribution of hosts across the IPv4 address space that can be broadly applied to any network monitoring or surveillance scenario requiring privacy preservation.
This page intentionally left blank.
Chapter 6
Conclusions and Future Research

6.1 Summary of Findings

This thesis has addressed the problem of estimating the malicious host population while preserving privacy. Based on our survey of the existing methodologies for estimating the malicious host population in Chapter 2 we identified four major open research problems that need to be addressed in order to find a better trade-off between the accuracy of estimation and the privacy of users. First was the lack of a model for host-to-address bindings. Second was the investigation of malicious address characteristics. Third was the correlation of independent measurements. And fourth was the development of dynamic countermeasures. We have addressed the first three of these open problems in Chapters 3, 4 and 5 respectively. The fourth problem remains to be addressed in the future.

The model of host-to-address bindings that we have proposed in Chapter 3 is probabilistic and enables the calculation of a conditional probability distribution of the number of hosts corresponding to an address given: a) the number of times the address has appeared during an observation window, and b) the latent probability of binding to a new host at each appearance. Based on reasonable assumptions of independence and stationarity, we developed parsimonious approximations of this model that are tractable and computable. We validated this approach using a four month trace of anonymised DHCP logs collected from a
Conclusions and Future Research

campus wireless network. Our major finding was the fact that hosts typically bind to addresses that appear more often, which is a manifestation of the phenomenon commonly known as preferential attachment. The major benefit of this model is that it allows inferring the number of hosts based purely on the characteristics of the addresses, which is a major requirement for privacy preservation that we identified earlier.

When it comes to actually applying our probabilistic model of host-to-address bindings, it is necessary to rigorously characterise the malicious address properties that can be observed by passive measurement techniques (e.g., intrusion detection systems and firewalls). To that end, in Chapter 4 we analysed a fortnight trace from DShield.org with aggregated intrusion alerts from over 1600 networks distributed around the world. We found that network intrusion alerts exhibit self-similar characteristics that lead to the loss of spatial and temporal locality. This results in characteristics like the overwhelming majority of addresses only appearing once during the measurement window. Furthermore, malicious activity originates from all parts of the IPv4 space in a broadly identical manner regardless of the category, volume and lifetime of addresses. We also found a diurnal cycle that is indicative of long-range dependence in terms of malware churn that leaves one address from an earlier interval and reappears as another address at a later interval. The implications of these characteristics is twofold. First, static blacklisting is bound to be ineffective over the long term due to such volatility and aliasing effects. Second, it is necessary to correlate intrusion alerts with more data in order to robustly infer the number of hidden hosts corresponding to each alerted address.

In Chapter 5, we combined intrusion alerts with ping responses as a way of alleviating the second implication resulting from the research contributions of the previous chapter. Our guiding hypothesis was that intrusion alerts and ping responses are different behavioural aspects of the same underlying malicious host.
6.2 Privacy Preservation

We tested and failed to reject this hypothesis on the basis of statistically compatible state transitions, sojourn intervals and long-range dependence found in both data sets. This allowed us to then unambiguously compile the intersecting time-series for each malicious address into alternating sequences of ON and OFF periods that made it possible to directly apply our probabilistic model of host-to-address bindings. Our major finding was that more than 80% of the addresses bound to more than one host and the distribution of malicious hosts across the IPv4 space was highly non-uniform. This has major implications for static blacklisting and the need for designing more dynamic countermeasures. Above all, we have demonstrated by addressing these three open problems that a better trade-off is possible in terms of the accuracy of estimation and the privacy of users. Figure 6.1 graphically summarises the better trade-off achieved by our methodology.

6.2 Privacy Preservation

The methodology developed in this thesis explicitly trades off the accuracy of estimation for the privacy of users. In particular, the definition of privacy preservation is that of decoupling the set of users corresponding to any given host from the instantaneous binding of that host to an address (Definition 1.5 in Chapter 1). This means that a user cannot be linked to an observed address with any kind of reliability even though we are able to make robust inferences about the set of hosts that attached to that address. In other words, the set of users and the associated set of hosts are fully hidden from the analyst in our methodology, while only exposing the observable characteristics of malicious addresses. This is a key contribution of the thesis and underpins the achievement of our primary goal corresponding to privacy preservation.
Figure 6.1: The relationship of our proposed methodology with respect to the current methodologies found in the literature when considering the critical requirements of accuracy and privacy.
6.3 Future Research

A number of future extensions is possible regarding the research presented in this thesis. We have identified four broad objectives for possible future extensions. First, there is scope for more extensive modelling of the host-to-address bindings. Second, there is always a need for newer data and ground truth data sets to be made publicly available so as to counter concept drift. Third, privacy issues in various jurisdictions need clarification for the research in this thesis to gain widespread adoption. Finally, dynamic countermeasures need to be developed based on the findings of this thesis. We elaborate further on these possible extensions next.

6.3.1 Modelling Extensions

One possible extension of our model of host-to-address binding is the addition of heterogeneity such that a mixture of the power-law and exponential approximations can be specified and tested. It is plausible that the power-law model is better suited to hosts that appear very infrequently, while the exponential model is better suited to hosts that appear more often. Hence, a hybrid approach that combines these two may be worthwhile in the future.

Another avenue worth pursuing is that of non-parametric Bayesian clustering [112]. These techniques work on the basis of an unbounded vector of parameters that only become defined by the available data, as opposed to the predefined parameterisation of our model. One potential benefit of applying such a clustering technique to the notion of alternating ON and OFF periods is that they might reveal those hosts that bind across multiple addresses in a way such that the overall population corresponding to a set of addresses could be inferred. It may also be worthwhile applying a form of “capture-recapture” analysis [130] for estimating the malicious host population on the Internet based on repeated sampling with
replacement (as suggested by [96] as well).

6.3.2 New Public Ground Truth Data Sets

We envisage a repository of anonymised data sets that provide various measurements of when a unique host bound to a unique address and for how long it stayed attached to that address. Such measurements can be carried out in a variety of settings, e.g., office and residential networks, wide area access networks, and so on. The benefit of having such variety of ground truth data sets would be that robust statistical characterisation of how hosts bind to addresses and vice versa could be carried out while still preserving privacy. As we have demonstrated in this thesis, such information can be used sensibly in order to counter malicious hosts while being non-intrusive and protective of the innocent users accidentally implicated in the activities of cyber-criminals.

6.3.3 Legal Precedence and Clarifications

Sharing and mediating access to sensitive data sets require legal support that is not uniform across different jurisdictions. Given that the requirement for privacy preservation in our work has been driven primarily by the strict Australian legislations [66, 65], it is necessary to seek out the precedence of any such legislation(s) in many other international jurisdictions. One of the obstacles to public access of ground truth data sets is the lack of clarity surrounding legal issues of privacy and confidentiality. What we envisage is the establishment of a “code of practice” for researchers in this specific area of network security and measurement that would define clear ethical and legal boundaries to stay within when sharing and mediating access to sensitive data sets.
6.3.4 Developing Dynamic Countermeasures

The inferences drawn using our method could be used to devise more effective countermeasures corresponding to the threat posed by malicious hosts. There might be possible uses in proactive blacklisting of addresses that are suspicious based on the number of hosts that have bound to them in the recent past. Further possibilities exist in terms of using such inferences to guide other classification and identification methodologies proposed in the literature, e.g., Ding et al. [32].
Bibliography


[14] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis, “My Botnet is Bigger than Yours (Maybe, Better than Yours): why size estimates remain chal-


Author/s: WAHID, ALIF

Title: Estimating the Internet malicious host population while preserving privacy

Date: 2013


Persistent Link: http://hdl.handle.net/11343/38139

File Description: Estimating the Internet malicious host population while preserving privacy

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.