Spatial-Temporal Characteristics of Internet Malicious Sources

Zesheng Chen
Florida International University
zchen@fiu.edu

Chuanyi Ji
Georgia Institute of Technology
jic@ece.gatech.edu

Paul Barford
University of Wisconsin-Madison
pb@cs.wisc.edu

***Abstract*—This paper presents a large scale longitudinal study of the spatial and temporal features of malicious source addresses. The basis of our study is a 402-day trace of over 7 billion Internet intrusion attempts provided by DShield.org, which includes 160 million unique source addresses. Specifically, we focus on spatial distributions and temporal characteristics of malicious sources. First, we find that one out of 27 hosts is potentially a scanning source among 2^{32} IPv4 addresses. We then show that malicious sources have a persistent, non-uniform spatial distribution. That is, more than 80% of the sources send packets from the same 20% of the IPv4 address space over time. We also find that 7.3% of malicious source addresses are unroutable, and that some source addresses are correlated. Next, we show that most sources have a short lifetime. 57.9% of the source addresses appear only once in the trace, and 90% of source addresses appear less than 5 times. These results have implications for both attacks and defenses.**

I. Introduction

With attack traffic on the rise, both defenders and attackers have a keen interest in identifying networks and hosts that are responsible for a significant portion of malicious activities or expose vulnerabilities. For example, attackers can implement a worm that propagates rapidly (*e.g.,* hitlist-scanning worms [11]) by focusing on hosts that are suspected to be vulnerable to specific exploits. On the other hand, defenders can use methods (*e.g.,* blacklisting [7]) to filter out the traffic from suspected malicious sources. Therefore, it is important to understand the locations of malicious sources in the IPv4 address space and how the locations evolve over time.

Recently, Casado *et al.* proposed the use of spurious traffic to gain insights on Internet measurements, such as the usage of NATs and the bandwidth of worm victims [4]. Inspired by their work, we analyze a large data set provided by DShield.org [13] to gain a deeper understanding of the behaviors and characteristics of sources that send malicious packets at Internet scale. DShield aggregates firewall and intrusion detection system logs from networks throughout the global Internet. Each log entry provided by a network represents one or more packets that violated a local rule. DShield transforms all of the logs into a normalized form. Each entry in the DShield trace includes: time-detected, submitter's ID, count, source IP, source port, destination IP, destination port, protocol-exploited, and flags. The source IP can be used for identifying a malicious/infected scanning source if the IP address is not spoofed, and is thus a focus of this work. Broadly speaking, the DShield trace provides a unique opportunity to extract the spatial-temporal characteristics of attacking machines.

In general, malicious sources are compromised hosts that can be either a part of IRC botnets or simply worm/virus victims. IRC bots operating in DoS mode send out requests to exhaust the resources on targets so that the targets cannot response to normal requests. On the other hand, worm infected systems constantly probe the Internet to find new vulnerable hosts. Compromised hosts are also used for network reconnaissance by scanning hosts to find available services or open ports. The traffic of these attacking sources can potentially be recorded by intrusion detection systems or firewalls. A subset of these records are submitted to DShield on a daily basis.

The task of protecting networks from scans and attacks has many significant challenges. First, there are no inherent security mechanisms in the current Internet architecture, and thus security must be implemented as a collection of "add-on" capabilities. Next, the sophistication of attackers and the code that they write is clearly on the rise [2], while the common user is often left behind from a technology standpoint - unable to decipher or keep up with good security practices. Finally, the increasing complexity and constant change and expansion in host and network technologies suggest that there will be new vulnerabilities (including zero-day exploits) in the foreseeable future. We believe that network-wide defenses, such as distributed firewalls [6], offer an opportunity to address these serious challenges, and thus understanding the details of source IP address characteristics is an important step toward making such systems viable.

In our prior work, a one-week DShield trace was used to show that malicious source addresses are highly unevenly distributed and form a relatively small number of tight clusters [1][5]. Moreover, this clustering feature can be characterized by a multiscale multiplicative innovations model and the Renyi information entropy. It is unclear, however, whether the highly uneven distribution of malicious sources is persistent over time. Furthermore, it would be interesting to understand the details of malicious source IP addresses in the small number of tight clusters.

In this paper, we extend the prior work to study the spatial-temporal characteristics of malicious sources in a greater detail and a larger scale. We focus on the two behaviors of malicious sources and attempt to answer the following questions:

- **Spatial behavior:** Is the distribution of malicious sources non-uniform across the IP address space over a long time

period? If so, which sub-networks are mainly responsible for malicious hosts? What are the invariants of the spatial distribution of malicious sources over time?

- **Temporal behavior:** How long are the lifetimes of malicious source addresses?

We also explore the correlations between the above spatial and temporal dimensions. We use the terminology "sources" to denote IPv4 source addresses in our data set.

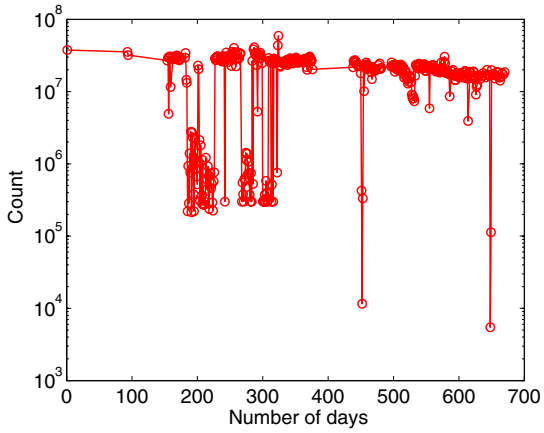
We use a 402-day trace from DShield.org that contains over 7 billion Internet intrusion attempts. We extract and analyze over 160 million unique source addresses from the trace. The size of this trace presents significant challenges in computation and analysis. The collected measurements form a unique, high-dimensional time series. There are more than 400 million hosts in the Internet today. If each host is a probable scanning source, the time series could have a dimension in the order of hundred-million! In fact, our first discovery is that 1 out of 27 host IP addresses is potentially a malicious scan source. Furthermore, malicious sources can send scan packets as frequently as tens per second. This can easily result in tens or hundreds of Gigabytes for a week's worth of data. These measurements exhibit complex and dynamic patterns as attacking sources are changing with time. All these factors make it a daunting task for identifying the features of malicious sources individually and in groups.

Our findings include the following observations: (a) Sources have a persistent non-uniform spatial distribution. More than 80% of sources concentrate on the same 20% of the IPv4 address space over time. Moreover, the top 20 prefixes contain 16.27% distinct sources. (b) Under the current state of network filtering, sources can still use IP spoofing techniques. For example, 7.3% of sources are unroutable that may result from IP spoofing. (c) Sources exhibit spatial correlations. For example, some top 20 prefixes belong to the same AS domain. (d) Most sources have a short lifetime. 57.9% of sources appear only in one day among the 402-day trace, whereas 90% of sources appear less than 5 times. Only few sources (0.04%) have a lifetime of no less than 100 days.

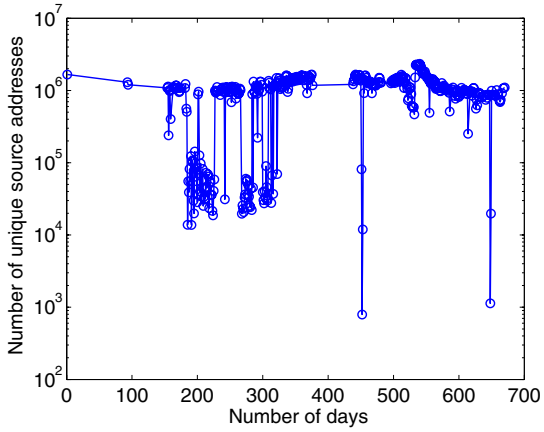
II. Data Set

Our data set is obtained from DShield [13]. DShield collects logs from firewalls and intrusion detection systems (IDS) from approximately 2,000 organizations distributed throughout the globe. The firewall/IDS platforms include BlackIce Defender, CISCO PIX Firewalls, ZoneAlarm, Linux IPchains, Portsentry, and Snort [12]. The logs are reported automatically by client programs running on the submitting hosts, typically once per hour. The objectives of the DShield are to provide aggregated information to the community, detect and analyze new worms and vulnerabilities, notify ISPs of exploited systems, publish the blacklists of worst offenders, and send feedback to submitters to improve firewall/IDS configurations. Our data set is a 402-day trace collected in the time period from November 10, 2004 to September 10, 2006[1]. This trace contains

[1]Among these 670 days, 402-day data are available to us.



(a) DShield records.



(b) Unique source addresses.

Fig. 1. Numbers of records and unique sources over time. The y-axis uses a *log* scale.

7,535,357,813 records. In these records, the total number of distinct source addresses is 160,590,790, which is in the order of 2^{27}. This suggests that one out of every 27 IP addresses in IPv4 could be considered as a potential malicious source. Hence, DShield observes a large number of IP addresses that are potentially IRC bots, worm infected machines, scanners, spoofed addresses, or other attacking hosts. A data set at such a large scale presents significant challenges in computation and analysis. For example, it took about a week for a computer with high power processors just to identify unique source addresses from the original data set.

Figures 1(a) and 1(b) show the numbers of records and sources observed on a daily basis. Note that the y-axis uses a *log* scale. The numbers of records are mostly in the order of 10^6 or 10^7, whereas for most days the numbers of sources remain in the order of 10^5 or 10^6. This implies that the appearance of sources is relatively consistent over time. It is noted that day 451 (Feb. 04, 2006) and day 647 (Aug. 19, 2006) have a lot fewer records and sources than other days and are likely to be caused by malfunctions in the DShield data collection system.

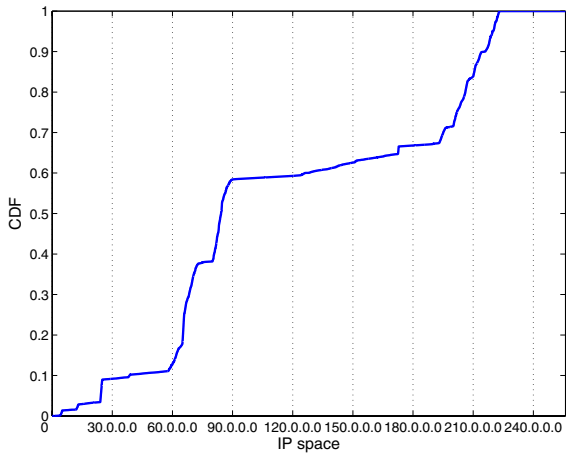


Fig. 2. CDF of sources of the entire trace in the IPv4 address space.

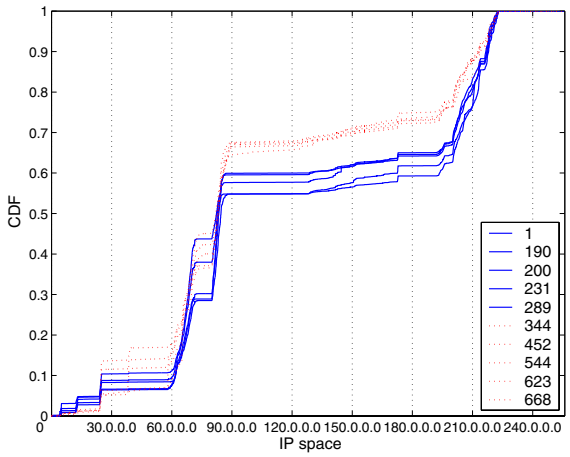


Fig. 3. CDF of sources in IPv4 address space for 10 randomly chosen days.

III. Spatial Behavior

In this section, we study the spatial behavior of sources. We start from aggregated observations and then delve into finer granularities.

1) Spatial Distribution: To quantify the distribution of sources in IPv4 address space, we consider the cumulative distribution function (CDF) of the distribution. Figure 2 shows the CDF of unique sources of the entire trace (160,590,790 addresses) in the IPv4 address space. Since the curve in some ranges of the IPv4 address space is much steeper than others, the distribution of sources is highly uneven. That is, some ranges of the address space contain a large number of sources. For example, 60.0.0.0/8 $\sim$ 90.0.0.0/8 includes almost 50% of sources, whereas 195.0.0.0/8 $\sim$ 220.0.0.0/8 holds about 35% of attacking hosts.

Next, we study whether the source distribution varies significantly over time, *i.e.,* individual days. To answer this question, we randomly select 10 days among 402 days and plot the distributions in these days in Figure 3. It can be seen that all curves in Figure 3 are similar to the curve in Figure 2 and do not vary significantly (similar characteristics were observed on other days). This implies that the uneven spatial pattern could be consistent across days.

2) 80-20 Rule: We further consider the uneven distribution characteristics of sources in /16 clusters. Specifically, we abstract 9 ranges of /16 subnets that account for a significant number of sources and summarize them in Table I. These subnets comprise 20% of all /16 subnets. Figure 4 plots the percentages of sources in these /16 clusters over time. We observe that except for day 451[2], the percentages are larger than 80%.

This observation is related to the well known *80-20 rule*. The 80-20 rule is also called the Pareto principle, the law of the vital few, or the principle of factor sparsity [16]. This rule states that for many phenomena, 80% of the consequences stem from 20% of the causes. We find that the distribution of sources is consistent with this rule: about 80% sources locate in the same 20% IP address space over time.

[2]We believe that this may be a outlier, since this day (Feb. 04, 2006) contains only 793 sources.

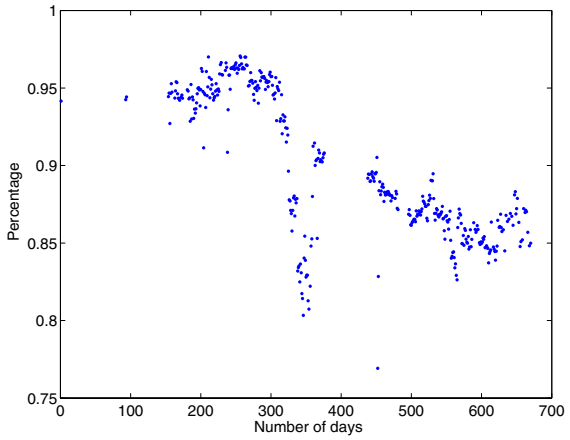


Fig. 4. Percentages of sources in /16 clusters over time.

TABLE I
9 ranges of /16 subnets or /16 clusters.

59.0.0.0/16 $\sim$ 72.57.0.0/16
80.0.0.0/16 $\sim$ 86.141.0.0/16
192.38.0.0/16 $\sim$ 196.204.0.0/16
198.53.0.0/16 $\sim$ 213.255.0.0/16
216.6.0.0/16 $\sim$ 222.253.0.0/16
4.0.0.0/8
12.0.0.0/8
24.0.0.0/8
172.0.0.0/8

3) Unroutable Sources: A domain may not be a /16 subnet. Thus, we continue our study by considering the source distribution among prefixes. Specifically, we extract a list of prefixes along with the corresponding ASes from the BGP repository at Oregon route-views [15] on December 19, 2006. The routing table contains 217,025 prefixes. We find that among 160,590,790 sources, 11,722,206 (7.3%) of the IP addresses do not match the corresponding prefixes, and are

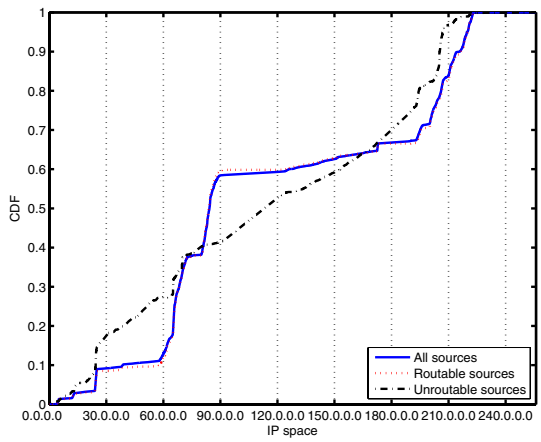
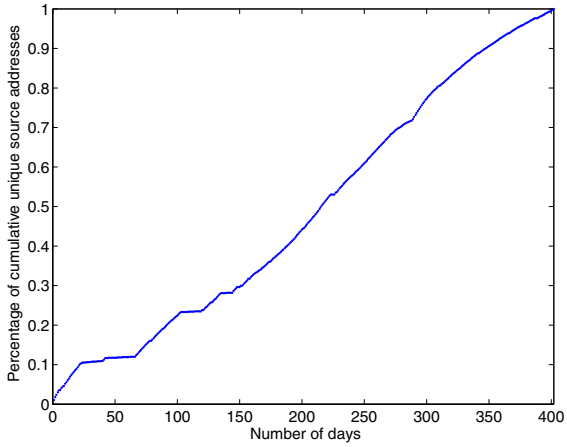


Fig. 5. CDF of unroutable sources in the IPv4 address space.

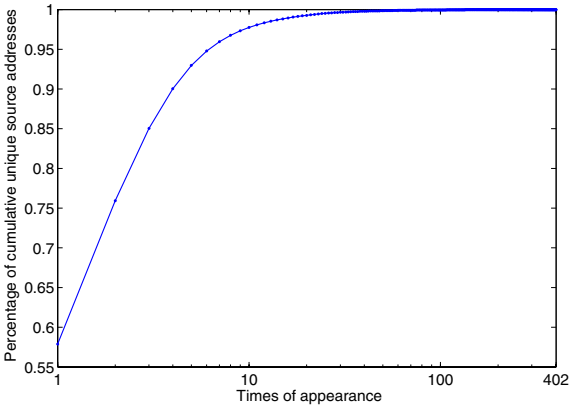
either truly unroutable or simply invisible from route-views. If they are unroutable, it implies that some attacks still use the IP spoofing techniques to hide the actual sources. We expect this to be the case. These could be DoS attacks or simply attacks for which a response is unnecessary. Figure 5 shows the distributions of all sources, routable sources, and unroutable sources. It is noted that the distributions of all sources and routable sources are almost identical, while unroutable sources are more evenly distributed than routable sources. The near-linear curve of the CDF of the unroutable-source distribution indicates that the spoofed addresses are relatively random. On the other hand, some regions of IPv4 addresses have more sources for both routable and unroutable sources, such as 65.0.0.0/8 $\sim$ 71.0.0.0/8 and 193.0.0.0/8 $\sim$ 220.0.0.0/8. One possible reason is that attackers used the spoofed IP addresses with the same short prefixes as the actual sources (like /8) to avoid the outbound traffic filtering.

Hence, the passive measurements from DShield indicate that IP spoofing techniques are still possible under the current state of network filtering and are used by attackers. This observation confirms the result in [3] based on active measurements.

4) Top 20 Prefixes: Based on the list of prefixes obtained from the route-views repository on December 19, 2006, we list the top 20 prefixes according to the number of sources in Table II. For each prefix, we also record the corresponding AS number, country, and ISP. The mapping from the IP prefix to ISP is based on the ip2location tool [14]. It is noted that the prefixes are not exclusive and some of them are overlapping, such as 4.0.0.0/8 and 4.0.0.0/9. There are totally 26,135,555 (16.27%) sources in these top 20 prefixes. We observe that among the top 20 prefixes, many of them are in the United States. Also, except 38.0.0.0/8, these prefixes belong to /16 clusters shown in Table I. Moreover, we are surprised to find that some top 20 prefixes belong to the same AS domain. For example, 84.128.0.0/10, 80.128.0.0/11, and 217.224.0.0/11 are in AS 3320. This demonstrates some degree of correlations among sources.



(a) Number of days.



(b) Number of appearances. The x-axis uses a *log* scale.

Fig. 6. Percentage of cumulative unique sources.

IV. Temporal Behavior

In this section, we study the lifetimes of all sources. We plot the percentage of cumulative unique sources for the entire trace in Figure 6. In Figure 6(a), the percentage of cumulative unique sources increases with time almost linearly. This indicates that each day DShield observes a large number of new sources that have not been recorded in the previous days. Figure 6(b) shows the number of appearances of unique sources. To our surprise, most of the individual IP addresses appear only a few times among 402 days. For example, 57.9% of sources appear only once, whereas 90% of sources appear less than 5 times. Only 0.04% sources have a lifetime of no less than 100 days. Thus, sources demonstrate significant, dynamic temporal behavior.

To further study the spatial-temporal behavior of sources, we divide sources into three groups according to the lifetime: a lifetime equal to 1; a lifetime in the range between 2 and 99; and a lifetime no less than 100. Figure 7 shows the spatial distributions of these three groups. It is observed that about 60% of sources are in the ranges of 60.0.0.0/8 $\sim$ 90.0.0.0/8 for the group that has a lifetime between 2 and 99, and 50% of sources are in the ranges of 195.0.0.0/8 $\sim$ 220.0.0.0/8 for the group that has a lifetime of no less than 100.

TABLE II
TOP 20 PREFIXES.

IP prefix	AS#	# of sources	Country	ISP
210.0.0.0/8	7474	3643064	AU	OPTUSONLINESERVICES-AP
84.128.0.0/10	3320	3115791	DE	DEUTSCHE TELEKOM AG
61.0.0.0/8	4678	2359184	IN	NATIONAL INTERNET BACKBONE
172.128.0.0/10	1668	2045237	US	AMERICA ONLINE
4.0.0.0/8	3356	1961772	US	LEVEL 3 COMMUNICATIONS INC
12.0.0.0/8	7018	1851807	US	ATT LINCROFT ORT
65.128.0.0/11	209	1781094	US	QWEST COMMUNICATIONS CORPORATION
65.192.0.0/11	701	1416188	US	UUNET TECHNOLOGIES INC
38.0.0.0/8	174	1031860	US	PERFORMANCE SYSTEMS INTERNATIONAL INC
80.128.0.0/11	3320	994636	DE	DEUTSCHE TELEKOM AG
83.0.0.0/11	5617	956003	PL	VOIP SERVICES BY POLISH TELECOM
82.224.0.0/11	12322	935487	FR	PROXAD / FREE SAS
65.112.0.0/12	209	881941	US	HARVARD UNIVERSITY
12.0.0.0/9	7018	760014	US	ATT LINCROFT ORT
86.128.0.0/10	2856	742166	US	BT-CENTRAL-PLUS
65.0.0.0/12	6389	737370	US	BELLSOUTH.NET INC
4.0.0.0/9	3356	598696	US	LEVEL 3 COMMUNICATIONS INC
81.128.0.0/11	2856	583225	UK	BT-N3SP
217.224.0.0/11	3320	567963	DE	DEUTSCHE TELEKOM AG
172.192.0.0/12	1668	530767	US	AMERICA ONLINE

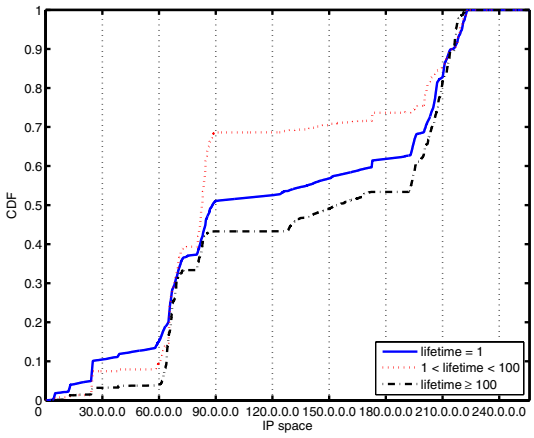


Fig. 7. Spatial distributions of sources with different lifetimes.

V. CONCLUSIONS

In this paper, we study the spatial-temporal characteristics of malicious sources at Internet. Our analysis is based on a huge trace provided by DShield.org that describes the long-term evolution of sources from a global viewpoint. We have focused on the distributional characteristics of sources in IPv4 address space and how actively a source sends malicious traffic in time. Our study leads to some interesting observations, which provide the implications on both attacks and defenses. For instance, 20% of the IP address space that contains over 80% of sources should be the focus for both attackers and defenders. Furthermore, the information on the spatial distribution of sources could be incorporated into the defense systems, such as packet filtering [8].

ACKNOWLEDGEMENTS

The authors would like to thank Johannes Ullrich and DShield.org for the data set used in this study. This work was supported in part by NSF grants CNS-0347252, CNS-0716460, ECS-0300605, and Cisco Systems. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or Cisco.

REFERENCES

[1] P. Barford, R. Nowak, R. Willett, and V. Yegneswaran, "Toward a model for sources of Internet background radiation," in *Proc. of the Passive and Active Measurement Conference (PAM'06)*, Australia, Mar. 2006.
[2] P. Barford and V. Yegneswaran, "An inside look at botnets," In Series: *Advances in Information Security*, Springer, Feb., 2007.
[3] R. Beverly and S. Bauer, "The spoofer project: inferring the extent of source address filtering on the Internet," in *Proc. of USENIX Steps to Reducing Unwanted Traffic on the Internet (SRUTI'05) Workshop*, Cambridge, MA, July 2005, pp. 53-59.
[4] M. Casado, T. Garfinkel, W. Cui, V. Paxson, and S. Savage, "Opportunistic measurement: extracting insight from spurious traffic," in *Fourth Workshop on Hot Topics in Networks (HotNets-IV)*, Nov. 2005.
[5] Z. Chen and C. Ji, "Measuring network-aware worm spreading ability," in *Proc. of INFOCOM'07*, Anchorage, AK, May 2007.
[6] S. Ioannidis, A. D. Keromytis, S. M. Bellovin, and J. M. Smith, "Implementing a distributed firewall," in *Proc. of the 7th ACM International Conference on Computer and Communications Security (CCS)*, Athens, Greece, Nov. 2000, pp. 190-199.
[7] D. Moore, C. Shannon, G. Voelker, and S. Savage, "Internet quarantine: requirements for containing self-propagating code," in *Proc. of INFOCOM'03*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1901-1910.
[8] K. Park and H. Lee, "On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets," in *Proc. of ACM SIGCOMM'01*, San Diego, CA, Aug. 2001, pp. 15-26.
[9] A. Ramachandran and N. Feamster, "Understanding the network-level behavior of spammers," in *Proc. of ACM SIGCOMM*, Italy, Sept. 2006.
[10] C. Shannon and D. Moore, "The spread of the Witty worm," *IEEE Security and Privacy*, vol. 2, no. 4, Jul-Aug 2004, pp. 46-50.
[11] S. Staniford, V. Paxson, and N. Weaver, "How to 0wn the Internet in your spare time," in *Proc. of the 11th USENIX Security Symposium (Security'02)*, San Francisco, CA, Aug. 2002.
[12] V. Yegneswaran, P. Barford, and J. Ullrich, "Internet intrusions: global characteristics and prevalence," in *Proc. of ACM SIGMETRICS'03*.
[13] Distributed Intrusion Detection System (DShield), http://www.dshield.org/.
[14] Ip2locaton, http://www.ip2location.com/.
[15] University of Oregon Route Views Project, Advanced Network Technology Center, University of Oregon, http://routeviews.org/.
[16] Wikipedia, "Pareto principle," http://en.wikipedia.org/wiki/Pareto_principle.