Correlation among piecewise unwanted traffic time series

Kensuke Fukuda
National Institute of Informatics
kensuke@nii.ac.jp
Tokyo, 101-8430, Japan

Toshio Hirotsu
Toyohashi University of Technology
hirotsu@ics.tut.ac.jp
Toyohashi, 441-8530, Japan

Osamu Akashi
NTT Network Innovation Labs.
akashi@core.ntt.co.jp
Tokyo, 180-8585, Japan

Toshiharu Sugawara
Waseda University
sugawara@waseda.jp
Tokyo, 169-8555, Japan

***Abstract*— In this paper, we investigate temporal and spatial correlations of time series of unwanted traffic (i.e., darknet or network telescope traffic) in order to estimate statistical behavior of unwanted activities from a small size of darknet address block. First, from the analysis of long-range dependency, we point out that TCP time series has a weak temporal correlation though UDP time series without huge flooding is well-modeled using a Poisson process. Next, we analyze the spatial correlation between two traffic time series divided by different sized darknet address blocks. We confirm that a TCP SYN traffic time series (e.g, virus or worm) has a clear spatial correlation in the arrival of packets between two neighboring address blocks. Indeed, this spatial correlation remains in traffic time series 1,000 addresses far from the target time series, even if a darknet address block is small (e.g., /26). On the other hand, TCP SYNACK traffic (e.g., backscatter) and UDP traffic (e.g., virus or worm) have less spatial correlation between two adjacent large address blocks. Finally, we estimate the average propagation delay of global unwanted activities appearing in TCP SYN traffic by using the generalized inter-correlation coefficient.**

I. Introduction

Nowadays, the Internet is one of the essential communication infrastructures for our daily life. However, compared to the old Internet, when only researchers accessed it, the current Internet requires much more attention to security issues. The detection and countervailing of Internet traffic anomalies (e.g., worms, virus, DDoS) is one of the hot topics in network research [2], [3], [7]–[9], [13], [16].

Currently, there are two types of approaches for detecting and characterizing traffic anomalies; active and passive methods. A well-known active approach is a honeypot/honeynet [15], which is a host that emulates a vulnerable operating system by using a virtualization technique. A honeypot host responds to a malicious activity so as to collecting a typical traffic pattern of such activities. The other approach uses a darknet (network telescope, internet motion sensor) [1], [8]–[10]; A darknet is a kind of blackhole of unwanted packets, and the corresponding hosts for that block do not actually exist although the address block for the darknet is advertised by a common routing protocol (i.e., BGP). Thus, the monitoring host easily obtains and identifies such unwanted packets that are mainly generated by software (e.g., scanners, worms, or viruses), backscatter, or misconfigured hosts.

However, there are two issues in the current passive approach of collecting unwanted packets for deploying to the current and future Internet. One is the limitation of IPv4 address blocks. Usually, in order to obtain the unwanted traffic accurately, one has to prepare a large address block (e.g., /8, /16). As recent arguments on the depletion of IPv4 addresses suggest [5], however, it will be more difficult to use such a large address block for monitoring in the future. The second issue is a scalability. All incoming packets to the darknet are only for initialization of connection or backscattered, because the monitoring host never responses to any incoming packets in the passive approach. However, the daily volume of unwanted packets without payload is a few giga-bytes for /18 sized darknet.

Thus, we need a sophisticated method for estimating unwanted traffic behaviors by observing only small address blocks (in which the number of incoming packets are relatively small). In this paper, we discuss temporal and spatial dependencies of the time series of unwanted traffic. In other words, the predictability of course-grained unwanted activities from smaller darknet address blocks is of our main interest.

II. Dataset

The dataset we analyzed in this paper was pcap packet traces captured at a darknet consisting of a consecutive /18-sided address block ($16,384$ addresses) from Oct. 2006 to May 2007. The mean and standard deviation of the number of packets in the dataset for one day is shown in Table I.

TABLE I
Mean and standard deviation of packet arrival [packets/day]

	mean	standard deviation
TCP SYN	604K	312K
TCP SYNACK	97K	100K
UDP	599K	391K

We focused on three types of time series of unwanted traffic reported by Ref. [10]. (1) TCP SYN: a time series was conformed by TCP packets only with a SYN flag representing the new connection requests. These packets are mainly generated by a network/port scanner, flooding, a virus, or a worm. (2) TCP SYNACK; The packets in this category have both SYN and ACK flags in the TCP header, and are known as backscatter (or background radiation) of (D)DoS victim, which is caused by an original packet with a spoofed source IP address. (3) UDP; Most of such packets are due to flooding, viruses, or worms. The main purpose of this study is to better understand the macroscopic behavior of unwanted traffic. Thus, we did not analyze a TCP/UDP time series categorized by the destination port number in order to avoid the explosion of the calculation cost, though such time series might be more informative. Also, we omitted ICMP packets since we found that those from other ASes were filtered at the border routers in the measurement AS.

Figure 1 indicates the fluctuation in the number of unwanted packets that arrived at our darknet address block (bin size: 1day). We can confirm that the fluctuations for the three categories are highly variable, but not correlated with each other.

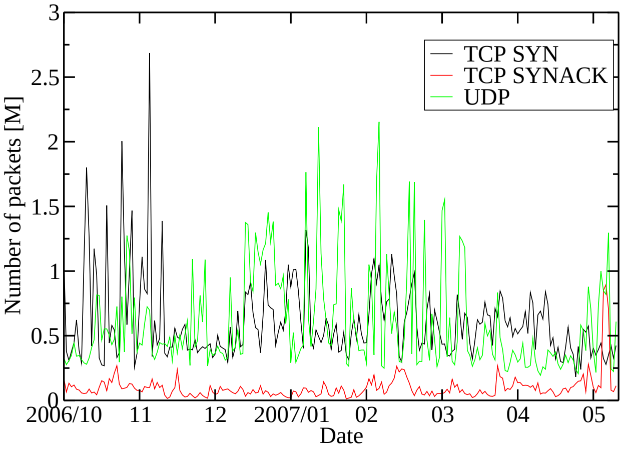

Fig. 1. Number of unwanted packets to darknet (size: /18)

III. Analysis Results

A. Temporal correlation of unwanted traffic time series

First, we focused on pieces of fixed-sized darknet IP address blocks (size: /k block) split from one large size darknet address block (size: /18). Thus, there were n sub address blocks $(0, 1, \cdots, n-1)$ with size /k. For each address block, we reconstructed a time series of unwanted traffic for two protocols (i.e., TCP and UDP) whose bin size was 1 min.

In order to characterize the temporal correlation of a traffic trace, we analyzed the long-range dependency (i.e., self-similarity) of each unwanted traffic time series per sub darknet address block. We evaluated the unwanted traffic time series by using the detrended fluctuation analysis (DFA) method [4], [12], though there are many methods for estimating the long-range dependency (LRD) of a given time series [11]. The DFA method of characterizing a nonstationary time series is based on the root mean square analysis of a random walk. An advantage of using the DFA method is that it produces results that are independent of the effect of the trend.

A detailed description of the DFA method is discussed by Ref [12], but here we will only briefly explain the method. We first integrate the time series, then divide it into "boxes" of length n. In each box, we calculate the least-squares polynominal fit of order p to the integrated signal (we used $p = 1$ in this study). Finally, in each box, we calculate the root-mean-square deviations of the integrated signal from the polynominal fit. We repeat the above procedure for different size boxes. For a LRD time series, we find the power law relation

$$F(n) \sim n^{\alpha}, \tag{1}$$

between the average magnitude of the root-mean-square deviations $F(n)$ and the size of boxes n.

The value of exponent α in Eq.(1) is the parameter that quantifies the statistical property of the time series; $\alpha = 0.5$ corresponds to the white noise, meaning that bursts in the time series do not correlate with each other. For $0.5 < \alpha \leq 1.0$, a burst in the time series is positively correlated in time, i.e. if one observes a burst, there is a high probability of observing similar sized bursts in the following time steps. Thus, current and future behavior depends on past events. On the other hand, $0 < \alpha < 0.5$ means that the time series has a negative correlation, indicating that a larger value will yield a smaller value, and vice versa. Importantly, α is closely related to the exponent of the power law β appeared in the power spectrum analysis; $\beta = 2\alpha - 1$, i.e., the Hurst parameter in an LRD time series directly corresponds to α.

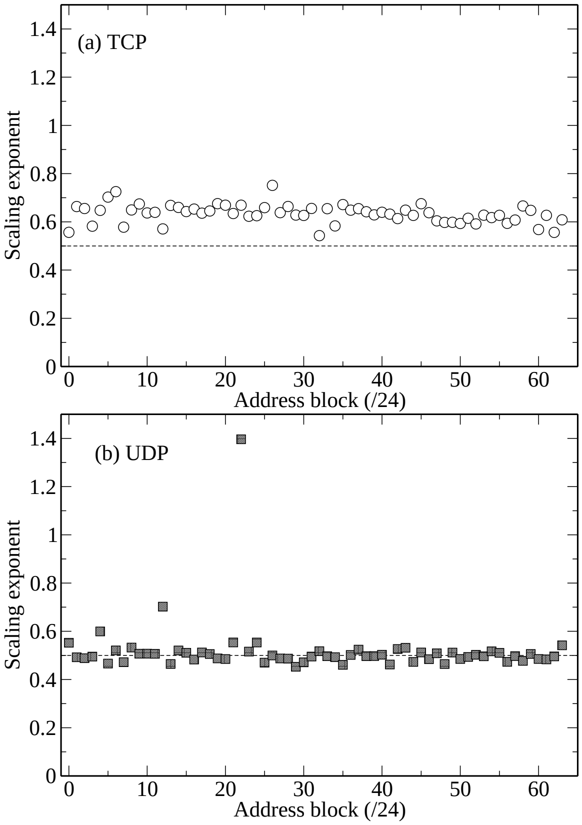

Fig. 2. Estimation of scaling exponent

Figure 2 indicates the estimation of the scaling exponent (α) of the power-law distribution for each /24 address block. For TCP traffic time series (upper figure), the exponents are scattered around 0.6-0.7, though they are clearly above 0.5; the traffic time series has a weak positive correlation. These exponent values for unwanted traffic, however, are smaller than those appeared in ordinary traffic time series (0.8-1.0) as reported by Ref. [11]. In our TCP traffic trace, most traffic volume was composed of connection initialization packets with a TCP SYN flag. Because the connection arrival time of general TCP flows are followed by exponential decay (i.e., short-range dependency), the smaller values of the scaling exponent we observed likely resemble those appeared in the connection arrival distribution. A plausible explanation of the weak positive correlation above 0.5 may be due to the retry of the TCP connection initialization [14].

On the other hand, for UDP traffic time series, the most values of estimated exponents were approximately 0.5, meaning that the traffic fluctuation also exhibits short-range dependency, and is close to white noise. One outlier whose exponent was 1.4 indicates clear non-stationarity, and we verified from manual inspection that this traffic trace contained a UDP flooding attack.

From both of these results, we concluded that the unwanted traffic time series has less temporal correlation, and it is not affected by a past unwanted activity. In addition, they imply the difficulty in temporal prediction of the anomalous activities from a time series observed at a single point.

B. Spatial correlation among IP address blocks

Next, we focused on the spatial correlation of time series among sub darknet address blocks. The purpose of analyzing the spatial correlation was to obtain knowledge on whether we can estimate the statistical behavior of unwanted traffic spread in the wider IP address block from a narrower darknet address block.

In order to characterize the correlation between two unwanted traffic time series, we calculated the correlation coefficient [6]. For given two time series ($T_\ell(t)$ and $T_m(t)$), the correlation coefficient $C(T_\ell, T_m)$ is defined as

$$C(T_\ell, T_m) = \frac{\sum(T_\ell(t_i) - E[T_\ell(t)])(T_m(t_i) - E[T_m(t)])}{\sqrt{V[T_\ell(t)]}\sqrt{V[T_m(t)]}}, \tag{2}$$

where $E[\cdot]$ and $V[\cdot]$ are the mean value and the variance of the time series, respectively. $C(T_\ell, T_m)$ characterizes the degree of similarity in the two time series as follows. $C = 0.0$ indicates that two time series are non-correlated. $0.0 < C \leq 1.0$ corresponds to a positive correlation, showing that both time series statistically resemble each other; by definition, $C = 1.0$ for $T_\ell(t_i) = T_m(t_i)$. Moreover, $-1.0 \leq C < 0$ indicates an anti-correlation, i.e., $T_\ell(t_i)$ takes a smaller value for larger $T_m(t_i)$, and vice versa.

Figure 3 displays the spatial correlation coefficient between two unwanted traffic time series for varying the size of sub darknet address block ($k = 23, 24, 25, 26, 27$). The x-axis represents the distance between two sub darknet address blocks denoted by the number of addresses. For example, the distance of $1,024$ addresses means that two sub darknet blocks are separate from 4 blocks in /24, or 8 blocks in /25. For TCP SYN packets (Fig3 (a)), we can confirm that the values of the correlation coefficient are larger than 0.4 for $\approx 1,024$ addresses, though a smaller address block result indicates a smaller value of the correlation. As for the largest value of the correlation, the value for /23 block (512 addresses) was around 0.6 at $1,024$ addresses. However, even for smaller address block (e.g., /27 block (32 addresses)), the correlation stayed around 0.4. Consequently, this result indicates that the shorter range of address blocks is still useful for estimating the statistical behavior of unwanted traffic for neighboring address blocks that are relatively far from the observed block.

On the other hand, for TCP SYNACK packets (Fig3(b)), the value of the correlation rapidly decreased even for the adjacent sub darknet address block. This means that we cannot expect any similarity of time series between adjacent address blocks. This was expected because these packets are mainly due to the backscatter of the DoS attack whose original packet had a spoofed source address.

Also, the results of UDP packets (Fig3(c)) were characterized by different behavior than the TCP SYN results. We again confirmed rapid decay of the correlation independent of the difference in the darknet sub address block size. The value of the correlation stayed around 0.25 at $1,024$ addresses, meaning that the arrival of UDP packets at an address block has less similarity to those at adjacent address blocks.

C. Estimation of average propagation delay of unwanted packets

In the previous subsection, we implicitly assumed that incoming relatedpackets from an unwanted activity to the wider darknet address block occurs within the bin size (1 min.). In other words, a scanning activity to a large number of hosts finishes within the bin size. However, if certain software (i.e., scanner, virus or worm) probe to a wider address range more slowly, the value of the correlation given in Eq.(2) cannot capture such behavior. Thus, we extend Eq.(2) to calculate the correlation between two time series with delay τ,

$$C(T_\ell, T_m(\tau)) = \frac{\sum(T_\ell(t_i) - E[T_\ell(t)])(T_m(t_i + \tau) - E[T_m(t)])}{\sqrt{V[T_\ell(t)]}\sqrt{V[T_m(t)]}}. \tag{3}$$

Equation (3) is known as the generalized inter-correlation coefficient [6]. We obtained the maximum value of $C(T_\ell, T_m(\tau))$ for $-3 \leq \tau \leq 3$. Such τ with maximum C corresponds to the average (or most significant) propagation delay between neighboring address blocks.

Figure 4 represents the estimated average propagation delays for two time series. TCP SYN (Fig.4 (a)) indicates stepwise plots per 1 min. delay, because the bin size of the time series was set to 1 min. However, it is clear that the procedure detects the significant delay of the two time series in minutes. From the plots, the average propagation delay of TCP SYN traffic time series was estimated as about $4,500$ addresses/min for 8 months. More interestingly, we could not observe distinguishable dependency of the size of sub darknet address blocks. Thus, in order to estimate the average

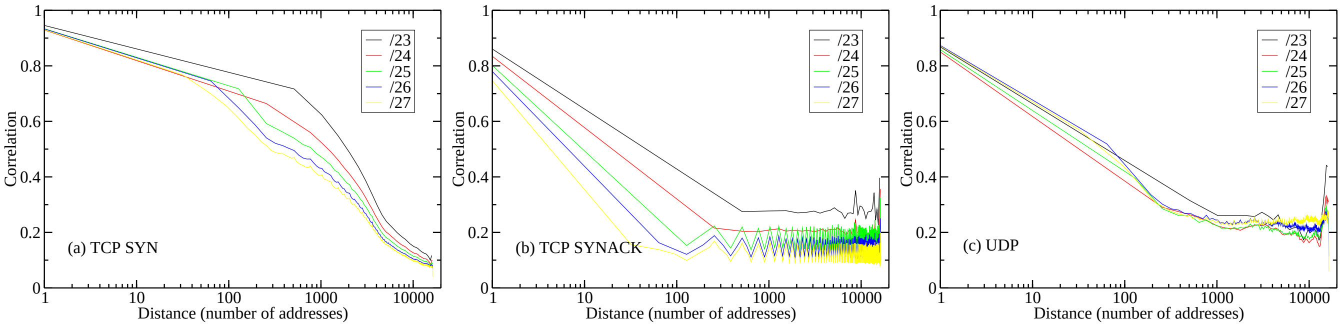

Fig. 3. Spatial correlation among sub darknet address blocks

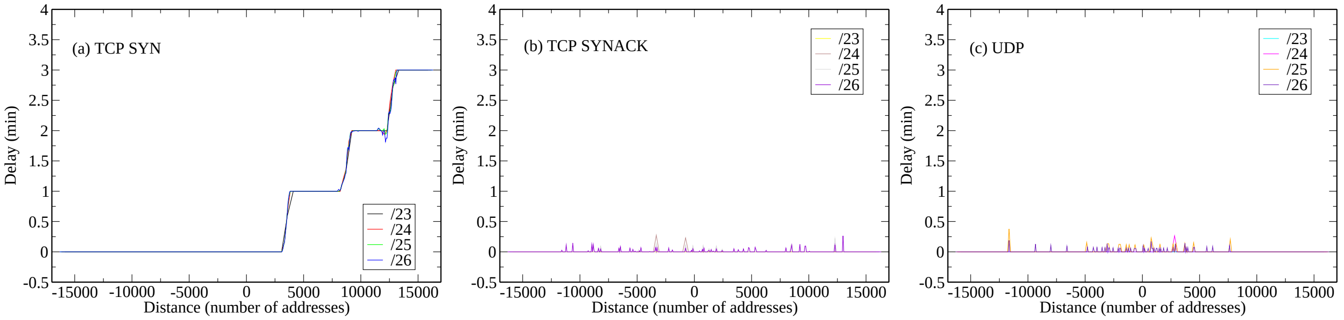

Fig. 4. Estimation of propagation delay

propagation delay of unwanted traffic time series, we only need small size of darknet address blocks (e.g, /26) in the current environment.

On other hand, for TCP SYNACK and UDP time series, there were no clear average propagation delays during the eight months of observation as shown in Figs.4 (b) and (c). Thus, those time series have no spatial correlation with adjacent time series, even considering the existence of the propagation delay. Consequently, we emphasize that the statistical behavior of TCP SYN traffic time series is intrinsic, which is contrary to that of the other time series.

D. Probability of the co-appearance of the same source IP address

Finally, we focused on how often the same source IP address appears in both neighboring sub darknet address blocks. We calculated the probability of the co-appearance of the same source IP address between two traffic traces of adjacent /24 address blocks. The bin size of the traffic trace was set to 1 hour. For each 1-hour data trace, the probability of co-appearance is given as follows;

$$p = \frac{1}{n-1} \sum_{i=0}^{n-2} \frac{|A_i \cap A_{i+1}|}{|A_i|}, \tag{4}$$

where A_i is a set of unique source IP addresses in the i-th sub darknet block.

As shown in Table II, for TCP SYN traces, the probability of the co-appearance was smaller than the other two types of traces. In the TCP SYN traces, there are a huge number of unique IP addresses, however, the number of packets from most IP addresses was only one or two. This is why the probability is relatively low. Moreover, by manual inspection, we confirmed that a few source IP addresses commonly appeared in the TCP SYN traces are source of many packets (i.e., port/address scan) to the darknet address block.

TABLE II

AVERAGE PROBABILITY OF THE APPEARANCE OF THE SAME SOURCE IP ADDRESS IN ADJACENT /24 DARKNET BLOCK (BIN SIZE 1 HOUR)

TCP SYN	TCP SYNACK	UDP
0.20	0.75	0.63

On the other hand, in the UDP traces, the probability of the appearance remained high, though the total number of packets arriving to the darknet was large. This is likely caused by the random probing mechanism for finding a target host.

IV. CONCLUDING REMARKS

In this paper, we discussed the temporal and spatial correlations among piecewise unwanted traffic time series to determine whether we can estimate statistical properties of global unwanted traffic behavior from smaller darknet address blocks.

We first showed that each unwanted time series split into sub darknet address blocks is characterized by a weak positive correlation for TCP packets, and no correlation for UDP packets from the DFA method. Thus, the fluctuation of unwanted traffic is close to random, compared to normal traffic. These results imply that it is statistically difficult to estimate the amount of current unwanted traffic from the traffic obtained in the past measurement.

Next, we pointed out that the TCP SYN traffic time series (e.g., virus or worm) has strong spatial correlation between neighboring observed sub darknet address blocks. This dependency decreases when the observed address block is far away from the target address block to be estimated. We have obtained that even a /26 address block (32 addresses) is enough to estimate the statistical behavior of address blocks far from about $1,024$ addresses. Thus, for TCP SYN behavior, we do not need to prepare a large address block for capturing the current dynamics of unwanted traffic. Moreover, we found that we can estimate the average propagation delay for significant activity of TCP SYN packet behavior by using a simple statistical method. We obtained the average propagation delay of 4500 addresses/min for eight months. This delay is much slower than the spread speed of outbreak of virus [13], so we should interpret this as the propagation delay of global unwanted activities. This estimated delay is a likely enough time scale for a network operator to manually write a new filter rule for border routers in order to filter out such unwanted traffic, after a report from another operator who manages a network with a lower address block in a 32-bit address space.

On the other hand, for TCP SYNACK (e.g., backscatter) and UDP (virus or worm) traffic time series, we confirmed that there is less correlation even between large adjacent neighboring sub address blocks (/23). Also, we could not estimate the average propagation delay of these packet behaviors, which is clearly different from the result of TCP SYN. Of course, the traffic generation mechanism for TCP SYNACK completely differs from the others, because of the background radiation of a (D)DoS attack. Consequently, the packet arrival for TCP SYNACK is likely well-modeled by a Poisson process in a wider range of the observed darknet address block.

Moreover, there are some arguments for UDP traffic. We expected that the basic behavior of UDP unwanted traffic resembles that of TCP SYN traffic, because we believe that the basic mechanism of a scanner, virus or worm is the same in both TCP and UDP. In reality, however, our findings showed that we should distinguish both of them; TCP SYN traffic still has some statistical structure in the coarse-grained time series, though UDP traffic has no temporal or spatial correlation at all.

We understand that our results are preliminary, and there are more points to be clarified. The main issue to be solved is to characterize a typical time scale to estimate the correlation structure from a given time series. In particular, the estimation of the average propagation delay strongly depends on the bin size and duration of the time series. We have obtained experimental results from weekly level time series, in which we estimated different (both faster and slower) average propagation delays in some cases, though there were no correlation structures in other cases. Moreover, currently we obtained only the propagation delay from a lower address in a 32-bit address space to a higher address. However, generally we cannot assume that the propagation delay will always be positive (i.e., lower to higher) in the future. Also, in order to accurately estimate the propagation delay, the bin size must be short, though there is a minimal bin size for statistical calculation. Thus, we need to carefully investigate further to determine the propagation delay, particularly for the time scale issue. Moreover, we need a more sophisticated classification for the UDP unwanted traffic. Clearly, one alternative is to use the information of the UDP destination port number, corresponding to the typical behavior of a worm or virus. It is likely a trade off problem between processing time and accuracy of identification.

ACKNOWLEDGMENTS

This research was partly supported by the Grant-in-Aid for Scientific research on priority areas (Info-plosion) from the Ministry of Education, culture, sports, science and technology (MEXT) Japan.

REFERENCES

[1] M. Bailey, E. Cooke, F. Jahanian, and D. Watson. The internet motion sensor: A distributed blackhole monitoring system. In *Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, Feb. 2005.

[2] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *SIGCOMM Internet Measurement Workshop (IMW2002)*, pages 71–82, Marseille, France, Nov. 2002.

[3] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho. Extracting hidden anomalies using sketch and non-gaussian multiresolution statistical detection procedures. In *SIGCOMM Workshop on Large Scale Attack Defense (LSAD2007)*, pages 145–152, Kyoto, Aug. 2007.

[4] K. Fukuda, L. A. N. Amaral, and H. E. Stanley. Dynamics of temporal correlation in daily internet traffic. In *IEEE Globecom2003*, pages 4069–4073, San Francisco, CA, Dec. 2003.

[5] G. Huston. Ipv4 address report. `http://www.potaroo.net/tools/ipv4/index.html`.

[6] R. Jain. *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, 1991.

[7] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *SIGCOMM'04*, pages 219–230, Portland OR, Aug. 2004.

[8] D. Moore, C. Shannon, D. Brown, G. M. Voelker, and S. Savage. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems*, 24(2):115–139, May 2006.

[9] D. Moore, C. Shannon, and J. Brown. Code-red: a case study on the spread and victims of an internet worm. In *SIGCOMM Internet Measurement Workshop (IMW2002)*, pages 273–284, Marseille, France, Nov. 2002.

[10] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of internet backgournd radiation. In *4th ACM SIGCOMM Conference on Internet Measurement (IMC2004)*, pages 27–40, Sicily, Italy, Oct. 2004.

[11] K. Park and W. Willinger. *Self-similar network traffic and performance evaluation*. John Wiley & Sons, 2000.

[12] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationry heartbeat time series. *Chaos*, 5:82–87, 1995.

[13] S. Staniford, D. Moore, V. Paxson, and N. Weaver. The top speed of flash worms. In *WORM2004*, pages 33–42, Washington, DC, Oct. 2004.

[14] W. R. Stevens. *TCP/IP Illustrated, Volume 1: The Protocol*. Addison-Wesley, 1994.

[15] The Honeynet Project. Know your enemy: Honeynets. `http://www.honeynet.org`, 2003.

[16] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan. Network anomography. In *SIGCOMM Internet Measurement Conference (IMC2005)*, pages 317–330, Berkeley, CA, Oct. 2005.