This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2010 proceedings

Estimating Speed of Scanning Activities with a Hough Transform

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

Romain Fontugne
The Graduate University for Advanced Studies
Tokyo, 101-8430, Japan
romain@nii.ac.jp

***Abstract*— In this paper, we propose a method to detect scanning activities in darknet traffic and to estimate their speed of change in time and feature space (e.g., destination address, source port, or destination port). The main idea of the algorithm relies on an image processing technique applied to a two-dimensional image that represents unwanted traffic. Thus, on the two-dimensional image, packets are represented as pixels in the time and feature coordinates, and unwanted activity as a set of pixels. The use of a Progressive Probabilistic Hough Transform (PPHT) that is a known technique to detect edges in an image enables us to detect such unwanted activities as "lines" in a traffic trace. We apply our method to darknet traffic traces for three years to investigate the property of such unwanted activities. Our main findings are following: In destination IP address space we confirmed typical host scanning speeds (i.e., a slanted line in the image) although the most of activities are characterized by intensive scans to a specific host (i.e., a horizontal line). Also, we confirmed few port scanning over wide destination port space, meaning that a targeted port attack is dominant in the current network. On the other hand, the consecutive change of source port was also observed; those activities are not tracked by other features. We obtain that 80-90% of unique source IP addresses appeared in the trace is confirmed by this method. Thus, most unwanted activities is still characterized by some kind of trajectory to be detected in packet feature space, though the rest of them behaves like "noise".**

I. Introduction

There have been much attention to network security in the Internet, since the Internet is one of the necessary infrastructures in our daily life. It is well-known the existence of unwanted (anomalous) traffic caused by virus, worm, misconfiguration, or DDoS passing through a backbone link, hidden in legitimate (normal) traffic. Characterization of such unwanted traffic is a hot topic in the current research field.

There are mainly two types of data measurement for analyzing unwanted traffic. The passive one is based on passive network sensor called "darknet" (or network telescope) that is a network address block in an edge network whose route is advertised to the Internet but has no normal hosts [2], [4]. Thus, all of the packets arriving at the darknet are naturally unwanted ones. The other one is semi-active measurement that deploys honeypots in an edge network; the host running honeypot is intensionally remained known security holes [20]. This can obtain more detailed information about the unwanted activities rather than the passive approach, because it reacts them so as to get their communication pattern. Thus, it is helpful in understanding microscopic behavior of the activities, however generally has drawback of the scalability.

In order to characterize unwanted traffic, many types of statistic-based approach have been applied to passive or semi-active data traces. In microscopic level, the characterization of specific worm activity has been well studied [16], [19] as well as the quantification of DDoS activity [17], [15]. Global unwanted activities is also analyzed in [21], [18], [1]. In the context of anomaly detection, finding hidden anomalies in backbone network has been a hot topic in network security [13], [5], though some statistical methods are only focused on volume-based anomaly [3], [12]. The detection of scanning activities based on a statistical test is also one of the interests [14], [10], [1], [7].

In this paper, we focus on a statistical characterization of unwanted traffics, especially their spreading speed; it is important to estimate and predict their impact, though there have been few result. We propose a method to estimate a scanning speed of unwanted activities in feature (i.e., address/port) space. The basic idea of the study is to analyze a packet as a pixel and a consecutive scanning activity in time as a "line" on a two-dimensional image (time and feature) by using edge detector based on a variant of the Hough transform called the Progressive Probabilistic Hough Transform (PPHT)) [9]. The Hough transform is a basic and well-studied edge detection algorithm in the image processing, and it is interpreted as a transform between the Cartesian coordinate plane and the polar coordinate plane. Thus, it naturally include the concept of the speed of activity in our context, because the polar coordinates are represented by radial and angular. Moreover, one of the advantages of the Hough transform is its robustness to random noises in a given image. Also the original line does not need to be strictly consecutive. To demonstrate the effectiveness of the use of the Hough transform, we report the result obtained by the proposed algorithm to 1-day darknet traffic traces measured at /18 darknet during 2006-2008.

II. Methodology

The goal of the proposed algorithm is to detect source hosts that are characterized by consecutive change of header features in packet in time. The algorithm consists of three parts; pre-processing, the PPHT, and post-processing. Here,

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

we first explain the basic and crucial part of the algorithm (the PPHT), then describe the rest of them.

A. Basic idea – the Progressive Probabilistic Hough Transform

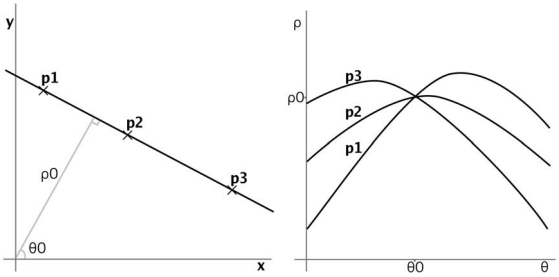

Fig. 1. Hough transform

The main idea of the algorithm relies on a basic image processing algorithm called the Hough transform [6], which is used for edge detection in an image. Let us consider a two-dimensional image representing network traffic whose x axis is the time bin a packet arrived at and y axis is one of the header features appeared in the packet header (e.g., destination IP address space). Each packet in the image is denoted by a pixel (x_i, y_i) in the Cartesian plane. The Hough (and reverse Hough) transform [6] is a transformation between the (two dimensional) Cartesian coordinate system and the (circular) polar coordinate system with majority vote. A point (x_i, y_i) in the Cartesian coordinate system holds a following relationship in the polar coordinate system; $\rho = x_i \cdot \cos\theta + y_i \cdot \sin\theta$, where ρ and θ are radial and angular coordinates, respectively (see also Figure 1). Now, a line in the Cartesian plane (the left figure) consists of three points (p1, p2, p3). A point is transformed into a corresponding trigonometrical curve in the polar plane (the right figure). The intersection point (θ_0, ρ_0) of the three curves in the polar plane represents the line in the original Cartesian plane. One can consider this transform as a majority vote in the polar plane (often referred as accumulator) by each input data in the Cartesian plane. In this example, the numbers of votes and that of majority vote are both three. A peaky intersection (i.e., it got many votes) is a candidate as a line to be detected. Conversely, the reverse Hough transform is given as $y = (-\frac{\cos\theta}{\sin\theta})x + (\frac{\rho}{\sin\theta})$, transforming an intersection point to the corresponding line in the original plane. Finally, one can reproduce the detected line by the Hough and reverse Hough transform. Additionally, if one can find such intersection points in the accumulator, the angular coordinate (θ) directly correspond to the speed (i.e., $\tan\theta$) of the scanning activity in our context.

Figure 2 depicts an example of the Hough transform in a darknet traffic trace. The upper figure is the original traffic data whose x-axis is the time and y-axis is the source port space. We can observe some slanted and horizontal lines in the figure by our eyeball. The middle figure is the corresponding representation in the polar plane, characterized by some clear

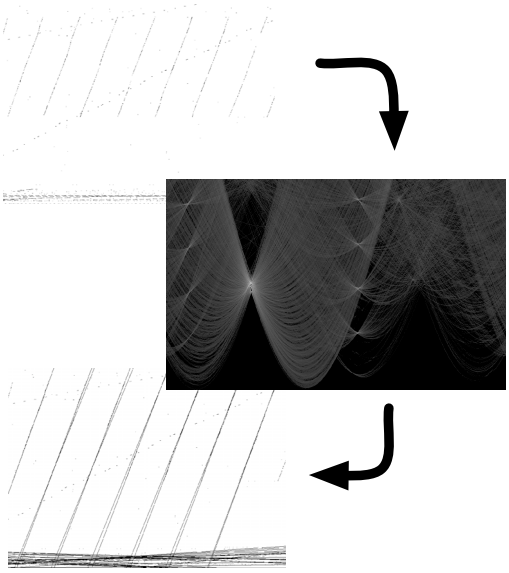

Fig. 2. Example of the Hough transform (upper: original darknet traffic (time vs source port), middle: the polar plane, bottom: detected lines

intersections of the curves. The bottom is the result of the reverse Hough transform applied to the significant intersections in the polar plane. We confirm that each line is successfully detected.

However, the original Hough transform has some drawbacks; the computation time is $O(n)$, where n is the number of pixels to be filled. Note that this is not equal to the number of packets in the trace. More seriously, it provides only θ and ρ, so that it is difficult to judge which point actually belongs to a detected line. To overwhelm these problems, a Progressive Probabilistic Hough transform (PPHT) [9] adopts random pickup of the pixels at the vote (for the reduction of n) and edge chasing by the knowledge of the derivation (θ) of the line at the vote (for obtaining the set of pixels belonging to a line). Finally, the PPHT enables us to provide a set of pixels belonging to a detected line as well as θ and ρ.

The outline of the PPHT is as follows [9]:

1) Check the input data, if it is empty then finish.
2) Update the accumulator with a single pixel randomly selected from the input data.
3) Remove pixel from the data
4) Check if the highest peak in the accumulator that was modified by the new pixel is higher than theshold (p_{th}). If not then goto 1
5) Look along a corridor specified by the peak in the accumulator, and find the longest segment of pixels either continuous or exhibiting a gap not exceeding a given threshold (ℓ_{th}).
6) Remove the pixels in the segment from the input data
7) Unvote from the accumulator all the pixels from the line that have previously voted.
8) If the line segment is longer than the minimum length, add it into the output list

9) goto 1

B. Feature

An image representing traffic consists of two dimensional coordinates; one is the arrival time of a packet (time bin) and the other is one of the features in packet header. A packet header contains many fields in IP/TCP(UDP), and generally all of them can be a candidate of the feature represented by the vertical axis in an image.

In this study, as a first step, we selected three typical header fields in a packet as the feature; destination IP address (DIP), destination port (DPORT) and source port (SPORT). DIP is the most basic feature to be analyzed. A detected horizontal and vertical lines correspond to an intensive probe to a target host and rapid host scan, respectively. A slow scan is detected as a slanted line with θ. DPORT is also an important field to identify the scanning activity. The last one, SPORT is not intuitive to be related to scanning activity since this field can be filled with an arbitrary value by a source host. However, it is suggested that consecutive change of this feature is sometimes useful in understanding unwanted traffic activity [7]; some source hosts re-use the same source port, but another uses an incremental value of SPORT.

C. Pre-processing

An input traffic trace was split into TCP and UDP packets, then each was aggregated into three (DIP, SPORT, DPORT) two dimensional time series (normalized by two values (0/1)). The size of time bin was set to 1s. The aggregation of DIP was performed by modulo operation (i.e., DIP (represented by 32bit space) % α, $\alpha = 8192$) in order to detect small change of the feature. Thus, a DIP is mapped into a value between 0 and 8192. On the other hand, for DPORT and SPORT, we used a normal aggregation, whose bin size was set to 8 ports (i.e., the number of bins was 8192). The size of the image for the calculation of the PPHT at once was 300 pixels (5 min) $\times$ 512 pixels. The parameters p_{th} and ℓ_{th} in the PPHT were empirically set to 10 and 6, respectively. Thus, the algorithm detects a line consisting of more than 10 packets in different pixels. Those setting were empirically determined, and a further study is needed for parameter tuning for other datasets.

D. Post-processing

After the PPHT, one obtains lists of pixels belonging to each line and estimated θ. The next step is to bind the pixels to corresponding source IP addresses. Note that one line is not necessary composed of a single source IP address; sometimes, a horizontal line for DIP or DPORT is caused by source hosts infected by the same virus. One calculates the median of θ for each source IP address in a figure as the representative θ, assuming that the host-level behavior in darknet is characterized by one specific origin (e.g., virus, worm). Finally, the detected IP address lists (DIP/SPORT/DPORT) are merged into a list of IP addresses.

E. Dataset

As a data set, we used three 1-day darknet traffic traces (pcap formart) collected at $/18$ $(16,384)$ address block in Japan during Nov. 2006 and Nov. 2008. The description of the traffic traces are shown in Table I. As previously noted, all of the packets appeared in the trace are unwanted one due to worm, virus, DDoS, or misconfiguration. The major destination ports were 4662, 445, 1433, 515, 143, 4669 for TCP and 7674, 1434, 137, 161, 53, 138 for UDP. Those are popular ports reported by network security vendors. However, a darknet packet is only the first packet in a flow, thus the exact identification of the cause is impossible. We eliminated ICMP packets in our evaluation because the uplink network of the darknet filtered all ICMP packets at the edges.

TABLE I
DARKNET TRAFFIC TRACES

trace	#tcp pkts	#udp pkts	#uniq IP (tcp)	#unique IP (udp)
Nov. 2006	598,968	392,890	37,130	38,225
Nov. 2007	802,305	480,826	11,723	8,427
Nov. 2008	1,283,044	728,132	91,776	18,059

III. RESULTS

A. Estimated speed of scanning activity

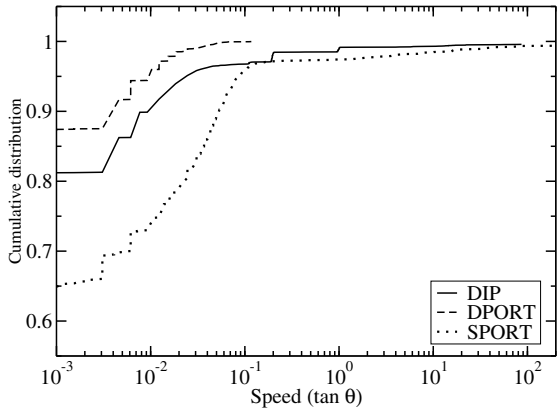

Fig. 3. The cumulative distribution of estimated speed of TCP scanning activity (Nov. 2006)

Figure 3 depicts the cumulative distribution of the estimated speed of scanning activity based on DIP, DPORT, SPORT for TCP packets on Nov. 2006. The x-axis represents the speed denoted by $\tan\theta$ (the unit is an IP address/sec for DIP, 8 ports/sec for SPORT and DPORT) and the y-axis is its cumulative distribution. A smaller value of speed corresponds to an intensive scanning to a fixed feature (i.e., a horizontal line) and a larger value does to a rapid change of the observed feature (i.e., a vertical line).

For DIP, the intensive scanning (horizontal line) accounted for 81% of the detected source hosts. It is a surprising result, considering the fact that data was taken from darknet; This address space has not been used. The ratio of the rapid host scan (vertical line) was only about 0.5%. 6% of them were

relatively slow scanning ($\tan\theta < 0.1$) , close to the intensive scanning. Similarly, we observed four characteristic jumps around $\tan\theta \approx 0.1$, 0.2, 1, and 15, representing typical speeds of scanning activity in the trace. Next, for DPORT, more than 87% of the source hosts correspond to intensive port scanning, meaning that source hosts look for hosts with known ports related to security hole. Interestingly, the fact we detected a few large-scale host scanning supports that the activity might be sophisticated to look for a target host. We could confirm a small portion ($< 0.01\%$) of port scanning over a large port space in this case. The curve of SPORT is different from others; only 65% of the source hosts used a fixed port. This is natural consequence that a source host lasts to use a fixed port from the initialization. However, about 10% of them were slowly changing its source port and 4% of them increased it rapidly. Furthermore, 0.6% of them correspond to instant changes (i.e., vertical line) of source port. The last consecutive behavior is likely a hint to find a hidden activity that is difficult to find by other features.

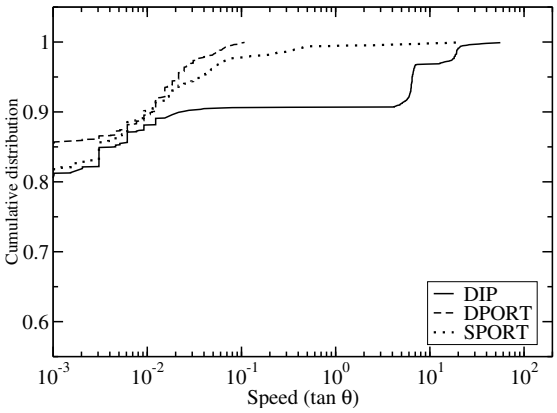

Fig. 4. The cumulative distribution of estimated speed of UDP scanning activity (Nov. 2006)

Figure 4 shows the same distribution for UDP packets. For DIP, we confirmed two characteristic big jumps around $\tan\theta \approx 8$ and 20. Both activities accounted for 9% of them, largely different from those for DIP of TCP. Also, about 80% of them were intensive scan (horizontal line), and a few host scans ($\approx 0.1\%$) were confirmed. As same as the case of TCP, there was few large-scale port scan in UDP, though over 85% of the source hosts scanned specific ports. Also, for SPORT, 80% of the source hosts used a fixed source port. We confirmed a small jump around $\tan\theta \approx 0.3$, indicating intrinsic behavior. 0.02% of them were characterized by rapid change of source port.

B. Difference in traffic traces

We have shown the snapshot behavior of scanning activities in the previous subsection. Next, we focus on a stability of the detected speed over time. Figure 5 and 6 are the cumulative distribution of scanning activities in 2007 and 2008, for comparison, respectively.

For TCP packets, the ratio of the intensive host scan decreased from 81% (2006) to 74% (2008). The typical jumps observed in Figure 3 were disappeared in 2007/2008 traces, though one jump was observed in $\tan\theta \approx 0.1$ for the recent traces. The ratio of the rapid scan increased to 1.3% in the latest trace. For SPORT, the shape of the curves were changed in time; there were two distinct jumps ($\tan\theta \approx 1$ and 4) in 2007, though those were not appeared in 2006. Furthermore, again two jumps were obscure in 2008.

For UDP packets, the ratio of the intensive host scan increased from 81% (2006) to 89% (2008). However, three characteristic speeds ($\tan\theta \approx 8$ and 20) of host scan were preserved, implying that the same type of programs (the same virus/worm or variants of them) had been active for the observed period.

Conclusively, the observed results pointed out the difficulty in modeling such activities by a simple numerical model.

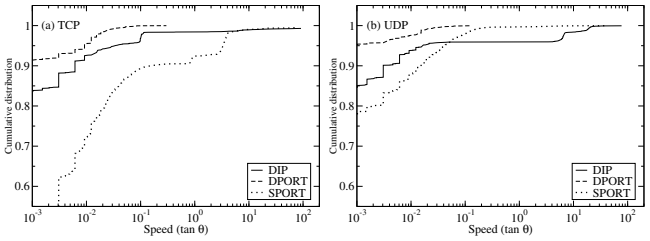

Fig. 5. The cumulative distribution of estimated speed of TCP/UDP scanning activity (Nov. 2007)

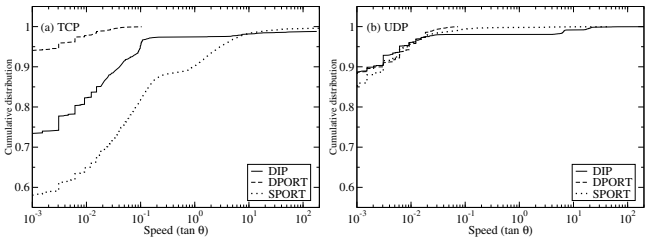

Fig. 6. The cumulative distribution of estimated speed of TCP/UDP scanning activity (Nov. 2008)

C. Coverage of the detected source IP address

In the previous subsection, we showed the results of estimated speed of unwanted activities. A natural question is raised on how many source IP addresses was detected in the trace. (i.e., the ratio of source hosts characterized by a line structure). Table II indicates the ratio of detected unique source IP addresses appeared in an image for the combination of three features. The ratio is denoted by the average over the images and its standard deviation.

Not surprisingly, the line structure by the three features was able to detect the source IP addresses appeared in the trace with high ratio. The ratio of the detected source IP addresses by our method (three features) to all the source IP addresses in the trace was 0.84 ± 0.04 for TCP and 0.90 ± 0.05 for UDP. This is mainly due to the effect of DPORT, exhibiting

TABLE II
COVERAGE OF UNIQUE SOURCE IP ADDRESSES IN THE TRACE (NOV. 2006)

	DIP	DPORT	SPORT	DIP ∪ DPORT	DIP ∪ SPORT	SPORT ∪ DPORT	ALL
TCP	0.58 ± 0.08	0.79 ± 0.04	0.06 ± 0.02	0.81 ± 0.05	0.62 ± 0.08	0.82 ± 0.04	0.84 ± 0.04
UDP	0.22 ± 0.30	0.89 ± 0.06	0.33 ± 0.23	0.90 ± 0.05	0.36 ± 0.25	0.89 ± 0.05	0.90 ± 0.05

horizontal lines. In other words, the other 10-15% is likely related to a flow with a sort of "random" fashion or a small flow consisting of a few packets.

The second point to note is that the ratios of DIP were relatively low (0.58 and 0.22). In particular, for UDP, 80% of the source IPs was represented without line structure. Thus, in the time-DIP plane, most of them looks like scattered "noise".

IV. CONCLUDING REMARKS

In this paper, we proposed a method to estimate the speed of scanning activities in darknet traffic. Our image processing approach is intuitive, however, it was possible to detect unwanted activities characterized by the consecutive change of features in the packet header field, though we confirmed 10-15% of random activities. Also, we demonstrated the time evolution of the speed of the unwanted activities during three years. The results showed the decrease of intensive host scan in TCP and the increase in UDP, and some specific speeds were preserved for the observed period.

There is a study focusing on the consecutive change of traffic feature to estimate the speed of unwanted activities[8], in which the speed is statistically estimated as the lag of two traffic time series of neighboring address block. However, the method has several drawbacks; it cannot detect intensive scanning (i.e., horizontal line), and only obtains one significant time lag between two neighboring time series. Also, it reported the difficulty in estimating the scanning speed in the UDP darknet traces. On the other hand, the PPHT successfully detected the anomalous behavior in darknet data, though the method is simple and intuitive.

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 time scale dependency of the proposed method. The parameter tuning is also a future work relating to this issue. Moreover, our algorithm focuses on the communication pattern of unwanted activities with temporal information. In this sense, our approach is similar to BLINC that is a graph-based traffic classification method [11]. One interesting possibility is to analyze an image consisting of two feature spaces (e.g., DIP and DPORT). Also, the application of our method to backbone traffic trace is important and realistic future work to characterize the scanning activity in the Internet. In a high-speed backbone network, the number of packets in a trace must be huge. In terms of the calculation cost, each image can be independently processed, so that the parallel computation will help us to get more processing performance.

ACKNOWLEDGMENTS

The authors are thankful to Toshio Hirotsu, Yosuke Himura, Patrice Abry, and Pierre Borgnat for their valuable comments.

REFERENCES

[1] M. Allman, V. Paxson, and J. Terrell. A brief history of scanning. In *ACM SIGCOMM IMC'07*, pages 77–82, San Diego, CA, 2007.

[2] 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.

[3] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *ACM SIGCOMM IMW'02*, pages 71–82, Marseille, France, 2002.

[4] E. Cooke, M. Bailey, Z. Mao, D. Watson, F. Jahanian, and D. McPherson. Toward understanding distributed blackhole placement. In *ACM WORM'04*, Washington D.C., 2004.

[5] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho. Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedure. In *ACM SIGCOMM LSAD'07*, pages 145–152, 2007.

[6] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[7] R. Fontugne, T. Hirotsu, and K. Fukuda. An image processing approach to traffic anomaly detection. In *AINTEC2008*, pages 17–26, Bangkok, Thailand, Nov 2008.

[8] K. Fukuda, T. Hirotsu, O. Akashi, and T. Sugawara. Correlation among piecewise unwanted traffic timeseries. In *IEEE GLOBECOM*, pages 1616–1620, New Orleans, LA, Dec 2008.

[9] G. Galambos, J. Matas, and J. Kittler. Progressive probabilistic hough transform for line detection. In *Computer Vision and Pattern Recognition*, pages 554–560, Los Alamitos, CA, 1999. IEEE.

[10] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In *IEEE Symposium on Security and Privacy*, pages 211–225, Oakland, CA, May 2004.

[11] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. Blinc: Multilevel traffic classification in the dark. In *ACM SIGCOMM'05*, pages 229–240, Philadelphia, PA, Aug 2005.

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

[13] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *ACM SIGCOMM'05*, pages 217–228, 2005.

[14] C. Leckie and R. Kotagiri. A probabilistic approach to detecting network scans. In *IEEE NOMS'02*, pages 369–372, Florence, Italy, Apr 2002.

[15] 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.

[16] D. Moore, C. Shannon, and J. Brown. Code-red: a case study on the spread and victims of an internet worm. In *ACM SIGCOMM IMW'02*, pages 273–284, Marseille, France, Nov. 2002.

[17] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of internet backgournd radiation. In *ACM SIGCOMM IMC'04*, pages 27–40, Sicily, Italy, Oct. 2004.

[18] A. Sridharan, T. Ye, and S. Bhattacharyya. Connectionless port scan detection on the backbone. In *Malware Workshop held in conjunction with IPCC*, Phoenix, AZ, Apr 2006.

[19] S. Staniford, D. Moore, V. Paxson, and N. Weaver. The top speed of flash worms. In *ACM WORM'04*, pages 33–42, Washington, DC, Oct. 2004.

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

[21] V. Yegneswaran, P. Barford, and J. Ullrich. Internet intrusions: Global characteristics and prevalence. In *ACM SIGMETRICS'03*, San Diego, CA, June 2003.