

Toward a Model for Source Addresses of Internet Background Radiation

Paul Barford, Rob Nowak, Rebecca Willett and Vinod Yegneswaran

University of Wisconsin and Duke University

E-mail: pb,vinod@cs.wisc.edu, nowak@cae.wisc.edu, willett@duke.edu

Abstract. Internet background radiation, the fundamentally unproductive traffic that arises from misconfigurations and malicious activities, is pervasive and has complex characteristics. Understanding the network locations of hosts that generate background radiation can be helpful in the development of new techniques aimed at reducing this unwanted traffic. This paper presents an initial investigation of the network locations of hosts that generate malicious background radiation using source addresses in packet traces from network telescopes, firewalls and intrusion detection systems distributed throughout the Internet. We characterize background radiation source addresses across the IPv4 address space for /8, /16 and /24 aggregates. Using a conservative multiscale density estimation method, we find that source addresses of background radiation form a relatively small number of tight clusters – *i.e.*, that the distribution of source addresses exhibits characteristics of a highly irregular multifractal with a broad spectrum that is consistent over all of our data. We verify that the distributional properties are consistent with multifractals, and propose a multiscale multiplicative innovations (MMI) model for host locations that can be used to generate random variates with the same distributional properties as our empirical data. This model is targeted for use in analytic, simulation and emulation evaluations of methods for reducing unwanted traffic as well as potential real time monitoring and detection applications.

1 Introduction

Internet background radiation is composed of the traffic generated by systems that are infected with malicious code and by systems that are either temporarily or permanently misconfigured (*e.g.*, [10]). The unwanted traffic generated by these systems is at best a relatively benign nuisance that consumes a modest amount of network and system resources. At worst, however, it represents a latent threat to all Internet resources, and is a harbinger of large-scale malicious activity. Recent work has shown that Internet background radiation has complex characteristics [9] suggesting that the long term objective of reducing or eliminating this traffic could be quite challenging.

It is likely that the potential impact of new techniques, tools and systems developed to reduce unwanted traffic will be evaluated analytically, in simulation or in emulation before being deployed. Thus, it is critical to have accurate models [5] for the basic mechanisms of background radiation such as exploit methods, targeting/scanning methods, and the corpus of hosts responsible for

this traffic. The objectives of our work are to provide an initial characterization study of malicious source addresses and to make a contribution to a set of models that can be used in general analysis of steps to reduce the *malicious* component background radiation.

The central question in our work is, where does Internet background radiation come from? To address that question, we conducted an empirical analysis of the *source addresses* of malicious Internet background radiation. The general problem of understanding characteristics of IP addresses in network packets was first addressed by Kohler *et al.* in [6]. That paper is concerned with understanding traffic aggregates through analysis of *destination* addresses in general IP traffic, and it serves as a guide for our work. Our study uses a large data set collected from network telescopes, firewalls and network intrusion detection systems distributed throughout the Internet. This multidimensional data set gives us a broad perspective on malicious background radiation traffic and enables issues such as spoofed source addresses to be treated, thereby enabling us to clarify the correspondence between source addresses and source locations.

We investigate source address distribution across the IPv4 address space in /8, /16 and /24 network aggregates. Despite the conservative nature of our analysis, our results show that the distribution of host locations is *not* smoothly varying. Instead, we observe a very bursty distribution, with a significant number of unique source addresses emanating from a small set of tightly concentrated network locations. The densities show remarkable structural consistency over each of our data sets, over time and for the different network aggregates. The complexity of the observed scaling structure suggests that it may be well modeled by a multifractal. We investigate this conjecture and find that indeed the distributions exhibit multifractal characteristics.

Based on our empirical evaluation, we propose a random cascade model to reproduce the observed scaling structure in source address densities. This model enables random variates to be generated with the same distributional properties as the observed data, and can be easily parameterized from source address data to reflect a wide range of behavior. Our model is parsimonious, only requiring estimation of a single parameter. Further, it easily accommodates unrouted or bogon regions in modeled address space enabling more realistic distributions. We provide details on variate generation and demonstrate that this technique generates statistically representative densities.

2 Data and Evaluation Approach

2.1 Empirical Datasets

Our empirical analysis is conducted on data sets collected from three different sources over a seven day period from 10-16 December 2004. The first is a large set of firewall and intrusion detection system logs provided by Dshield.org [16]. Dshield aggregates logs on a daily basis from over 1600 networks distributed throughout the globe. The logs provide a uniform condensed summary of malicious activity observed in each provider's network. The benefit of this data set

for our study is its broad coverage and the fact that source addresses are not obfuscated. The drawback is that we do not have ability to distinguish spoofed source addresses in this data set.

The other sources of data are two dark (unused) address space monitors that run the *iSink* system [19]. The first includes unused portions of two class B networks in use at our campus (referred to as “Campus” – approximately 16,000 addresses monitored) and the second is an entire class A network (referred to as “Provider” – 16 million addresses monitored). *iSink*’s active responder capability enables us to separate verifiably non-spoofed sources from the set of all sources by simply considering TCP sources that engage in complete TCP connection setup. The details of all three data sets used in our study is given in Table 1.

Name	#Scans	#Unique Source Addresses
Dshield	296,004,577	7,694,291
Campus	36,774,175	448,894
Provider	212,481,885	2,355,150

Table 1. Details of the data sets used in this study which were collected from Dec 10-16 2004.

We consider a single week’s data from all three networks and partition the dataset both across time and targeted service port. For each dataset, we compute and examine the distribution of the number of unique source hosts from each IPv4 source network. While most of the analysis is performed at the /24 source network granularity, we also consider modeling sources in /16 and /8 aggregates. We focus on ports 80 (HTTP), 135 (DCE/RPC), and 445 (NetBIOS/SMB) because these were services that were most actively targeted during that week.

2.2 Estimation of Source Address Distribution

The objective of our evaluation is to understand the distributional characteristics of background radiation source addresses throughout the IPv4 address space. This problem is challenging for several reasons, including the fact that host systems themselves are not uniformly distributed throughout the address space [6, 7] and the fact that background radiation is quite dynamic [2, 9]. We address the former issue by using the spatially adaptive multiscale analysis method described in the following section. We address the latter by breaking down the data sets described in Table 1 into daily aggregates and by scans to specific ports.

Our initial observations of source locations indicated a very bursty, spatially inhomogeneous distribution. This implies that characterizing the distribution accurately from measured data would require a spatially adaptive density estimation technique. Since standard estimators such as histograms do not adapt to spatial changes in the structure of the data, their density estimates are in practice frequently over-smoothed where the density is changing rapidly or under-smoothed where the density is changing more slowly. Such estimators do not preserve singularities or sharp changes in the underlying density. Furthermore,

selection of the appropriate histogram bin-width is a challenging problem and can introduce a significant bias to the estimator.

To overcome this challenge, we use a novel multiscale density estimation technique based on wavelet-type methods [18]. This method optimally adapts to unknown characteristics, ranging from highly localized spikes and abrupt changes to smooth variations. The density estimate can be computed efficiently in $O(n)$ time, where n is the number of detected source addresses, as follows. The estimation procedure computes an optimal partition of the address space and a local polynomial is fit to the data within the partition interval. The optimal partition is obtained via a tree pruning process, in which pruning corresponds to a merging of neighboring partition intervals. The pruning is data driven, enabling the procedure to adapt to variations in the activity in different source address regions. The polynomial fits within each partition interval are computed via a maximum-likelihood procedure. For more details on the theoretical properties of the estimation procedure and its implementation, see [18].

3 Analysis and Modeling

3.1 Evaluation of Source Address Distributions

We begin our analysis by considering the density of source addresses in our three data sets grouped into /8, /16 and /24 aggregates. The results are shown in Figure 1. If source addresses were uniformly distributed, our analysis would show smooth lines across the entire IPv4 address space. What we see, even at /8 aggregates, are sharp discontinuities, and sparsity with these effects increasing for /16 (not shown due to space constraints) and /24 aggregates. The figure shows striking similarity across the data sets at each aggregation level. This is a crucial result since the Campus and Provider graphs are for *non-spoofed sources* while DSHIELD data includes spoofed sources (we also verified this with the Campus and Provider data sets that included spoofed sources). The implication is that spoofing does not have a significant impact on the density of malicious source addresses. Since the /24 aggregate shows the highest level of detail, we will focus subsequent analysis at this level.

Our second analysis considers the temporal characteristics of malicious source addresses in each of the data sets. We generated density plots for each data set broken down into daily aggregates. The results for three of the days in each of the three data sets are shown in Figure 2. The figure shows that while source densities vary somewhat across networks, they are very consistent on a daily basis. The results of the temporal breakdowns (not shown due to space constraints) on other days were quite similar.

Our final analysis considers malicious source address density broken down by destination port. Figure 3 shows the density plots for each of the target services described in Section 2 for a single day aggregated by /24 source networks in all three data sets. The figure shows that while source address densities change across ports, the essential characteristics of sparsity and burstiness are maintained across the three data sets.

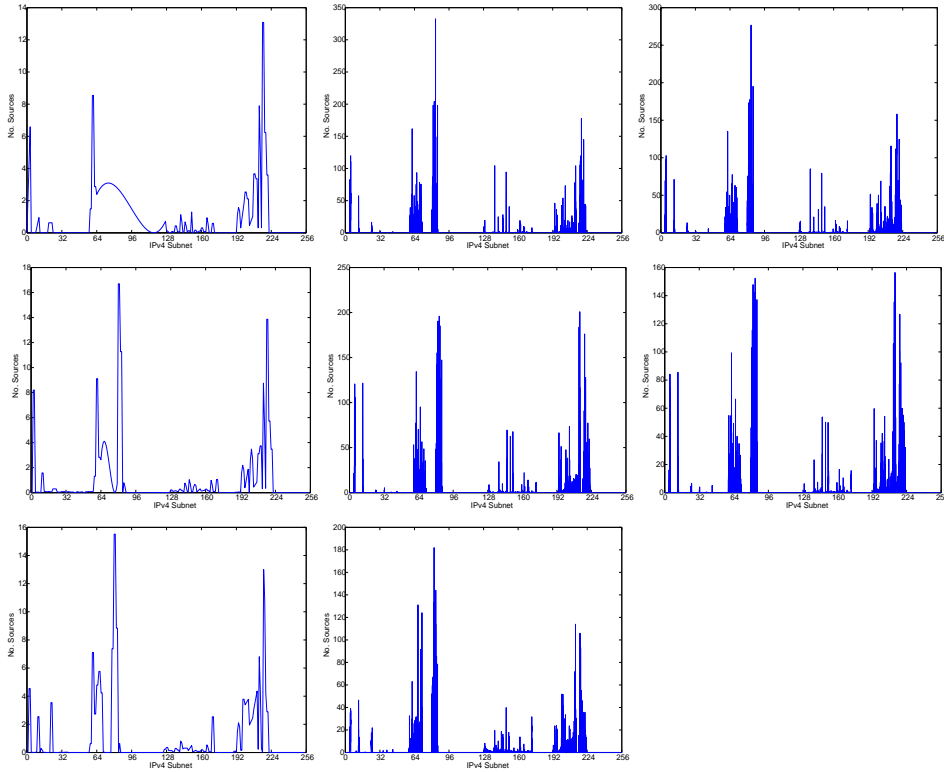


Fig. 1. Density plots for Campus (top), Provider (middle) and DSIIELD (bottom row) at /8 (left column), /24 (middle column) and /24-all (spoofed + non-spoofed) (right column). We do not have ability to obtain non-spoofed TCP DSIIELD data. Left and middle columns for Campus and Provider sites only include verifiably non-spoofed TCP source addresses. X-axis: Subnet, Y-axis: No. observed source addresses

3.2 A Random Cascade Model for Source Address Distributions

Kohler *et al.* [6] demonstrate that the distributions of destination IP addresses seen on Internet links have multifractal characteristics. The complex characteristics of the background radiation source address distributions led to our conjecture that their scaling structure may also be consistent with a multifractal. The left plot in Figure 4 shows a sample distribution of source addresses from a one day snapshot of the Campus data set (/24 aggregate) and on the right is its *multifractal spectrum*, which we calculated using the histogram method described in [6]. As in [6], sampling effects dominate our analysis at finer scales, however, analysis at medium scales show that the spectrum covers a wide range of values which is consistent with multifractal scaling (similar results were found for other data sets but are not shown due to space constraints).

This leads us to propose a simple algorithm for generating random IP distributions with multifractal characteristics similar to those of our empirical data. The model we propose is a random cascade model (also called a multiplica-

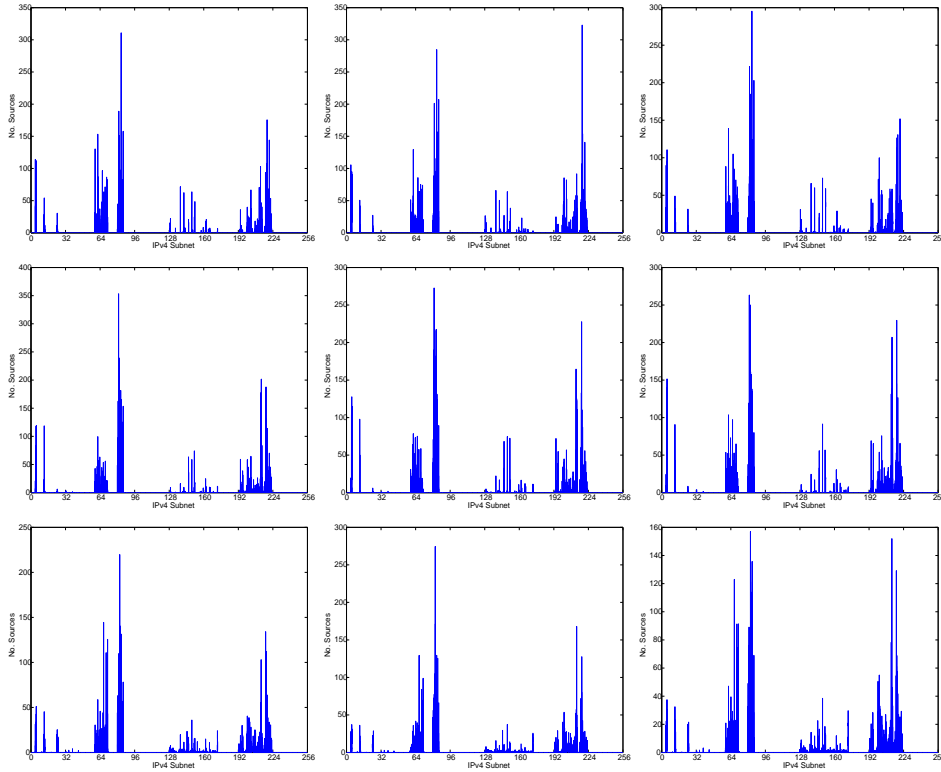


Fig. 2. Density plots for the Campus(top), Provider(middle) and DSCHILD(bottom) data set in /24 aggregates for three days (left-right: Dec 11, 13, 15 2004) of the measurement period. X-axis: Subnet, Y-axis: No. observed source addresses

tive, multiscale innovations (MMI) model [8]) that is similar in some respects to the model in [4]. We generate a density function defined on the interval $[0, 1]$ and then map this directly to the source IP address space, easily incorporating important Internet-specific characteristics such as unrouted /8 networks and bogon regions. Like most generative models for multifractals, ours begins by assigning unit mass to the entire interval $[0, 1]$, and then proceeds by recursively subdividing the interval and re-allocating the mass between the resulting subintervals. Unlike the “Cantor dust” model proposed in [6] which re-allocates mass according to a deterministic (Cantor set) rule, our model re-allocates mass in a probabilistic manner based on a parametric density function that can be easily fitted to the characteristics of the observed IP address data. The added flexibility in our model allows one to (randomly) generate new and distinct IP address distributions, all with multifractal characteristics matched to the observed data. This capability has obvious potential applications in simulation studies and testing. The Cantor dust model proposed in [6] is used to simply illustrate that its multifractal spectral characteristics can resemble those of IP address distributions, supporting the hypothesis that IP address distributions are multifractal

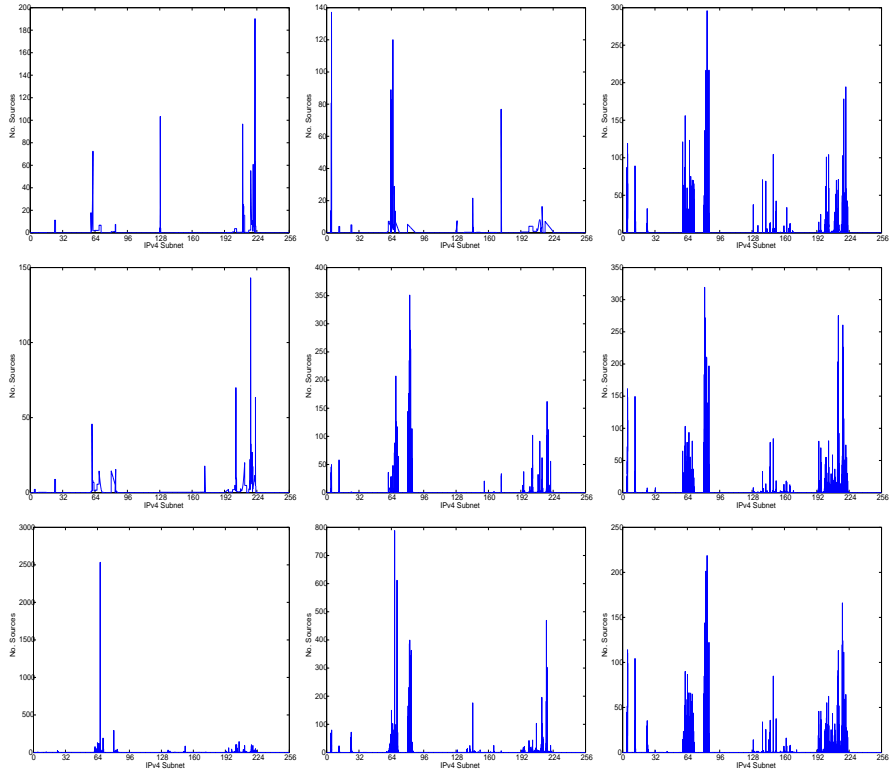


Fig. 3. Density plots for the Campus(top), Provider(middle) and DSCHILD(bottom) data set in /24 aggregates on a single day (Dec 11) broken down by port (port 80, 135 and 445 left-right). X-axis: Subnet, Y-axis: No. observed source addresses

in nature. Since the Cantor dust model is deterministic, it cannot be used to generate *different* IP address distributions for simulation purposes.

Our model can easily be described in terms of the scaling coefficients associated with the Haar wavelet transform. Let f denote an IP address density function. The (un-normalized) Haar scaling coefficient of f at scale 2^{-j} and location k is defined as the integral of f over the interval $[k2^{-j}, (k+1)2^{-j}]$ and we will denote it by $c_{j,k}$. A generative model for a density f is given by the following iteration in terms of the scaling coefficients. Begin with $c_{0,0} = 1$ and iterate according to:

$$\begin{aligned} c_{j+1,2k} &= c_{j,k} \rho_{j,k} \\ c_{j+1,2k+1} &= c_{j,k} (1 - \rho_{j,k}) \end{aligned}$$

for $k = 0 \dots 2^j$ and where the parameters $0 \leq \rho_{j,k} \leq 1$. This process can be repeated up to a sufficiently large value of j (e.g., $j = 24$ corresponds to /24 IP address resolution). The parameters $\{\rho_{j,k}\}$ control the re-allocation of mass between the resulting (dyadic) subintervals. The parameters can be deterministic (as in the Cantor dust model [6]) or they can be random variables.

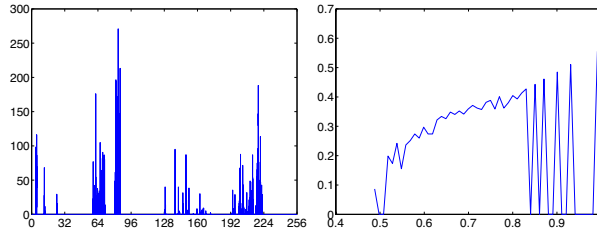


Fig. 4. Sample distribution of observed source addresses from the Dec. 10 (/24 aggregate) snapshot of the Campus data set (left) and its multifractal spectra (scale $j = 16$, Y-axis: scaling exponent).

Probabilistic models for the parameters allow us to not only model one specific IP address distribution, but also to generate new, random distributions with similar characteristics. In modeling the distribution of the ρ parameters, it is important not to fit the model to observations at the finest scale because the ρ 's observed here reflect the statistics of measuring a series of discrete occurrences rather than the structure of the underlying multifractal model. As a result, most of the observed ρ 's at the finest scales are factors of $1/2$, $1/3$, $1/4$, etc. Similar sampling issues were noted in [6]. We model the $\rho_{j,k}$ as beta distributed random variables. The beta density is appropriate in this context for several reasons. First, it is supported on the interval $[0, 1]$, as are the ρ 's. Second, it is a two-parameter density which allows for significant modeling flexibility. Let $\mathcal{B}e(\beta, \beta')$ denote the density. The mean of a random variable with this distribution is $\beta/(\beta + \beta')$, and so we typically assume the parameters $\beta = \beta'$, reflecting our prior assumptions about the regularity of the refinement process in the address space. As β tends to ∞ , our generative model tends to a uniform distribution over IP space. As β tends to zero, our generative model tends to a highly multifractal distribution similar to the Cantor dust model [6].

The multiplicative cascade model described above is well-known to possess multifractal characteristics [8]. In fact, its multifractal spectrum is given by:

$$S(\alpha) = \inf_{q \in \mathbb{R}} (q\alpha - \tau(q)),$$

where

$$\tau(q) = 1 - \log_2 \left(\frac{\Gamma(\beta + q)\Gamma(2\beta)}{\Gamma(2\beta + q)\Gamma(\beta)} \right).$$

The multifractal spectrum broadens as β decreases and narrows as β increases. Thus, the parameter β directly controls the multifractality of the distribution.

The model can be easily fitted to our data by simply computing the Haar scaling coefficients and solving for the parameters $\rho_{j,k}$ from them. By viewing the collection $\{\rho_{j,k}\}$ of the computed parameters as independent and identically distributed samples from an underlying beta distribution, we can easily find the maximum likelihood estimate for the best β parameter for the beta density; MATLAB has a simple function in their statistics toolbox called `betafit` for just this purpose. We did this using our data at scales eight and coarser (to

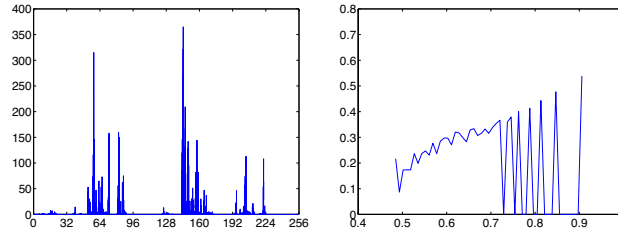


Fig. 5. Sample distribution of generated source addresses using parameters extracted from the distribution shown in Figure 4 (left) and its multifractal spectrum (scale $j = 16$, Y-axis: scaling exponent).

mitigate sampling effects as described above), obtaining a beta density parameter of $\beta = 0.61$ for the data set shown in Figure 4. Using this parameter we generated new source IP distributions which share the same multifractal characteristics as observed in the data. An example along with its multifractal spectrum is shown in Figure 5 (note, we did not embed unallocated address space information in this distribution).

4 Discussion and Conclusions

The essence of our empirical results is that source addresses of Internet background radiation are sparse, bursty and show consistent structure across different network monitors and over time. We consider these results preliminary due to the limited data set we consider and the documented dynamics of malicious traffic in general and background radiation in particular [2, 9]. We are tracking malicious source addresses on an on-going basis for this reason. Likewise, while our multifractal cascade model captures essential characteristics of source address density, we will continue to validate its robustness versus empirical results over time.

An important question that we will investigate more deeply in the future is what are the mechanism that cause malicious source addresses distributions to appear multifractal? The simple intuition offered in [6] is appealing place to start. Namely, that the IPv4 address space has been historically allocated as series of successively smaller prefix blocks that follow typically a left-to-right allocation within a given block. This process is virtually identical to our cascade model for multifractal generation. Further, the occurrence of infected hosts within an allocated address block is likely to be either zero (*e.g.*, in well managed networks), random (*e.g.*, ISPs with home users) or complete (*e.g.*, in poorly managed networks). Thus, the multifractal structure of malicious source addresses is most likely to be a direct reflection of multifractal structure of the overall IPv4 address allocation.

Our model for malicious source addresses is an initial contribution to what we hope will eventually be a library of useful models that characterize unwanted traffic behavior. There are many potential applications for such models including use in simulators such as ns2 [14] and large emulation testbeds such as [3, 15,

17]. For example, models of malicious code propagation (*e.g.*, [1]) coupled with models of wide area network topology *e.g.* at the autonomous system level [13] could be used with our model to evaluate how background radiation propagates throughout the Internet and/or how filtering techniques (*e.g.*, [11, 20]) could be employed to reduce this traffic.

Our analysis and model have more general implications if one assumes that most background radiation traffic is generated by systems that have been compromised. This implies that these machines are (*i*) currently vulnerable, (*ii*) probable targets for future worm outbreaks, and (*iii*) most likely to be harvested as zombies for launching future denial-of-service attacks. Thus, our source location analysis should be valuable for studying a broader range of malicious threats. For example, we are currently investigating the possibility of using changes in source address distributions as the basis for fast and accurate detection of new outbreaks and attacks.

Acknowledgements: The authors would like to thank Johannes Ullrich for the DSHIELD data logs and Dave Plonka and Geoff Horne for iSink support. This work is supported in part by NSF grant numbers CNS-0347252, ANI-0335234 and CCR-0325653. 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.

References

1. Z. Chen, L. Gao, and K. Kwiat. Modeling the spread of active worms. In *Proceedings of IEEE INFOCOM '03*, March 2003.
2. E. Cooke, M. Bailey, M. Mao, D. Watson, F. Jahanian, and D. McPherson. Toward understanding distributed blackhole placement. In *Proceedings of CCS Workshop on Rapid Malcode (WORM '04)*, October 2004.
3. DETER: A Laboratory for Security Research. <http://www.isi.edu/deter>, 2005.
4. A. Feldmann, A. Gilbert, and W. Willinger. Data Networks as Cascades: Investigating the Multifractal Nature of Internet WAN Traffic. In *Proceedings of ACM SIGCOMM*, 1998.
5. S. Floyd and E. Kohler. Internet research needs better models. In *Proceedings of the First Workshop on Hot Topics in Networks*, October 2002.
6. E. Kohler, J. Li, V. Paxson, and S. Shenker. Observed structure of addresses in ip traffic. In *Proceedings of ACM Internet Measurement Conference*, November 2004.
7. S. McCreary and K. Claffy. <http://www.caida.org/outreach/resources/learn/ipv4space>, 1998.
8. R. Nowak. Fractal modeling and analysis of poisson processes. In *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems and Computers*, November 1998.
9. R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of internet background radiation. In *Proceedings of ACM Internet Measurement Conference*, 2004.
10. D. Plonka. Flawed routers flood university of wisconsin time server. UW Tech Report, 2003.
11. P. Porras, L. Briesemeister, K. Skinner, K. Levitt, J. Rowe, and A. Ting. A hybrid quarantine defense. In *Proceedings of CCS Workshop on Rapid Malcode (WORM '04)*, October 2004.
12. R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Transactions on Information Theory*, 45(4), 1999.
13. L. Subramanian, S. Agarwal, J. Rexford, and R. Katz. Characterizing the internet hierarchy from multiple vantage points. In *Proceedings of IEEE INFOCOM '02*, June 2002.
14. The ns2 Network Simulator Project. <http://www.isi.edu/nsnam/ns>, 2005.
15. The Wisconsin Advanced Internet Laboratory. <http://wail.cs.wisc.edu>, 2005.
16. J. Ullrich. Dshield. <http://www.dshield.org>, 2005.
17. B. White, J. Lepreau, L. Stoller, R. Ricci, S. Guruprasad, M. Newbold, M. Hibler, C. Barb, and A. Joglekar. An integrated experimental environment for distributed systems and networks. In *Proceedings of OSDI*, December 2002.
18. R. Willett and R. Nowak. Multiscale poisson intensity and density estimation. submitted to *IEEE Transactions on Information Theory*. Available at <http://www.ece.rice.edu/willett/Research/publications.html>.
19. V. Yegneswaran, P. Barford, and D. Plonka. On the design and use of internet sinks for network abuse monitoring. In *Proceedings of RAID '04*, September 2004.
20. C. Zou, W. Gong, and D. Towsley. Worm propagation modeling and analysis under dynamic quarantine defense. In *Proceedings of CCS Workshop on Rapid Malcode (WORM '03)*, October 2003.