Opportunistic Measurement: Extracting Insight from Spurious Traffic

Martin Casado and Tal Garfinkel, Stanford University
Weidong Cui, UC Berkeley
Vern Paxson, ICSI
Stefan Savage, UC San Diego

***Abstract*— While network measurement *techniques* are continually improving, representative network *measurements* are increasingly scarce. The issue is fundamentally one of access: either the points of interest are hidden, are unwilling, or are sufficiently many that representative analysis is daunting if not unattainable. In particular, much of the Internet's modern growth, in both size and complexity, is "protected" by NAT and firewall technologies that preclude the use of traditional measurement techniques. Thus, while we can see the shrinking visible portion of the Internet with ever-greater fidelity, the majority of the Internet remains invisible. We argue for a new approach to illuminate these hidden regions of the Internet: *opportunistic* measurement that leverages sources of "spurious" network traffic such as worms, misconfigurations, spam floods, and malicious automated scans. We identify a number of such sources and demonstrate their potential to provide measurement data at a far greater scale and scope than modern research sources. Most importantly, these sources provide insight into portions of the network unseen using traditional measurement approaches. Finally, we discuss the challenges of *bias* and *noise* that accompany any use of spurious network traffic.**

I. Introduction

Much of our insight into the current state of the Internet derives from empirical measurement studies. Unfortunately, while the measurement techniques used in these studies are increasingly refined, the *scope* at which researchers can conduct such measurements is conversely shrinking.

For example, the growth of network address translation (NAT) has hamstrung traditional active measurement efforts - which typically presuppose addressibility. Thus even simple questions about edge network demographics are difficult to answer because researchers lack adequate *access* to the types of machines (i.e., home users, small businesses) that heavily determine the answer. Moreover, researchers are also limited to using *well-behaved* network traffic in their active measurement studies. It would be unthinkable to conduct a large-scale measurement study of bisection bandwidth capacities by flooding the network from thousands of sources. Indeed, the increase in network-borne threats has fueled a backlash against even the most innocuous network probes—a ping packet to many hosts produces a nasty e-mail in addition to a round-trip time measurement.

Consequently, much active measurement research relies upon dedicated infrastructures (PlanetLab, NIMI) to provide data. However, such infrastructures are inherently limited as the number of available sources are relatively small and homogeneous (e.g., 10s or hundreds of nodes associated with educational or research networks, often close to the core) and not representative of the larger Internet (millions of end-hosts in homes, small businesses, Internet cafes, often deep on the edge). Cooperative efforts to gain greater access to these sources [10], [2] have yet to see much adoption.

Similarly, passive measurement efforts are gated by the richness of the observer's vantage point. Most researchers are consequently limited to research and educational networks and typically a limited set of the links in those settings. Obtaining traces for a large and diverse demographic requires greater cooperation from network carriers, which, for business reasons and privacy concerns, has generally been infeasible. Exceptions to this rule have been performed by the carriers themselves [9], [7] and even these still only cover a tiny fraction of Internet hosts and paths.

Consequently, while current measurement techniques can tell us more than ever before about the visible portions of the Internet, those portions of the Internet with little or no visibility remain far larger, with the gap between visible and dark likely widening.

Addressing these problems would seemingly require widespread deployment of measurement agents *inside* edge networks, generating regular test traffic of sufficient scale and diversity to drive general experimentation. While a straightforward implementation of this vision is both economically and socially infeasible, in this paper we argue that such traffic is already being generated and can be opportunistically measured. In particular, we propose exploiting the prodigious, yet underutilized, traffic generated by compromised or misconfigured hosts—worm probes, botnet scans, DDoS backscatter, spam floods and so forth. In preliminary experiments we show that such data provides a broad, diverse and viable substrate for a variety of network measurement activities and serendipitously side-steps many of the limitations of traditional methods.

In the remainder of this paper we present our case in more depth. In the next section we discuss how spurious traffic can provide improvements in scale and diversity over traditional methods, followed by examples of large scale vents that generate such traffic in § III. In § IV we explore techniques for utilizing these sources and § V sketches some preliminary results from opportunistic measurement. We sketch the new challenges and limitations presented by this approach in § VI and finally we conclude in § VII.

II. Why Spurious Traffic?

Spurious traffic provides us with a number of unique characteristics that are attractive for network measurement.

1) Many Sources: Harnessing spurious events for measurement purposes can yield several orders of magnitude more traffic sources than are currently available from other active measurement sources such as PlanetLab [4] or neti@home [10]. Organized activity such as automated scanning and spam can use large bot networks of tens of thousands of hosts. For example, we have recorded traffic from over 16,000 unique IP addresses from a single concerted scan. In another study of a domain that receives large amounts of spam we recorded 38,000 addresses.

Internet-scale events such as worm outbreaks can infect hundreds of thousands of hosts, creating an incredibly large and topologically diverse pool of traffic sources. For example, CAIDA recorded traffic from 359,000 sources for the first Code Red (CRv2) outbreak [17] and 160,000 sources for NIMDA [1]. While the peak number of traffic sources from initial infections is relatively short lived, often large numbers of infected machines continue to generate useful traffic, sometimes years after the initial infection.

Moreover, most of the sources we have examined are distinct between spurious traffic episodes. For example, comparing sets of addresses from three major traffic sources—a large, automated scan ($\sim$ 16,000 machines), long-lived Code Red II infections ($\sim$ 1,500 machines) and hosts sending email to a heavily spammed domain ($\sim$ 38,000 machines), we find only 24 addresses that appeared in more than one set. Thus, we can potentially combine multiple types of sources to obtain even larger pools of sources, if the measurement we wish to perform is compatible with the different source types.

2) Great Diversity: Today's organized measurement infrastructures are highly homogeneous, consisting primarily of sources from academic institutions in the US and Western Europe interconnected with high-bandwidth low-congestion links. For example, PlanetLab currently has 584 nodes representing 275 sites in a handful of countries.

In contrast, sources of spurious traffic are often Internet-wide and biased towards machines that have been difficult for researchers to access (e.g., those of private institutions and individuals). For example, the 38,000 machines originating spam to one of our domains have `whois`records with addresses from 159 countries, and bottleneck bandwidth (measured using the M&M tool suite [11]) ranging from 56 Kbps to 622 Mbps. In addition, a large fraction of the machines we have measured reside behind NAT boxes (see § V). These hosts and their networks would be invisible to traditional measurement techniques.

3) Social Acceptability: Historically, it has been taboo to consider measurement activities that would consume very large amounts of aggregate bandwidth: generating wide-scale high-volume network probes is at best considered anti-social and irresponsible, and at worse as no different than a hostile network attack. However, measuring preexisting sources that exhibit such behavior (such as the Slammer worm's saturation of network access links [16]) raises no such concerns—the event has *already happened* due to someone else's misbehavior and there is no direct harm caused by exploiting this behavior.

Similarly, large-scale passive analysis of legitimate traffic raises very significant privacy concerns, and thus has been largely infeasible to date. However, spurious traffic is generally devoid of normal application content, rendering these issues moot. Furthermore, the unsolicited broadcast nature of this traffic provides some safe harbor even should the contents be sensitive or proprietary. In the vernacular, "the cat is already out of the bag".

III. Examples of Traffic Sources for Opportunistic Measurement

Opportunistic measurement requires discovering events (traffic sources) that satisfy several constraints. First, the traffic must meet the requirements of the analysis that we wish to perform: e.g., high-volume TCP flows for bottleneck bandwidth estimation, or long-term predictable traffic for path characterization. The event must also generate enough traffic to produce statistically meaningful results. Finally, the traffic must include destination addresses visible to the researcher. Regarding this last point, often spurious traffic events generate traffic viewable from any vantage point on the Internet. But in addition such traffic frequently has a non-uniform distribution, creating *hot-spots* or even *attractors* where the traffic concentrates.

We have observed several different classes of events exhibiting these properties:

A. Event Classes

1) Worms: Worms turn large numbers of hosts into traffic sources, and their code is directly available. These features make them ideal candidates for opportunistic measurement. Worms also provide two different modes useful for measurement. The initial flurry of traffic from a worm outbreak typically lasts for only a few hours, but results in traffic from a massive number of sources — sometimes numbering in the tens of thousands of machines. We refer to these singular events as *supernovas*. Taking best advantage of these spectacular measurement events requires careful prior planning to ensure that we have the necessary passive measurement infrastructure in place to capture the occurrence.

However, infected machines can continue to scan for longer periods, sometimes even years, after the initial attack. This ongoing activity can provide predictable long-term traffic sources that we term *pulsars*. Indeed, worms released in 2001 [17], [1] continue to scan the Internet from thousands of infected hosts.[1]

2) Automated Scans: Another significant source of traffic on the Internet is the ever-present "background radiation" of automated scans by attackers looking for vulnerable machines [20]. Malicious scans are often performed collaboratively by large collections of bots, sometimes numbering in the 10,000s of machines. Their scanning patterns vary widely from unpredictable sharp bursts to slow linear probes lasting weeks.

Automated scans differ from worms in that generally they seek to derive more information about a host. Where a worm is typically interested only in finding the next target, a tool-driven scan may look for specifics of a given protocol stack, multiple vulnerabilities, or other remotely discernible information. Consequently, network scans can generate relatively large amounts of traffic to individual IP addresses (or subnets) within a short time period.

3) Spam: When present in large quantities, spam provides an interesting class of spurious traffic because it gives us access to relatively long-lived TCP flows. Passive measurement tools often require ample flow sizes (e.g., 50 packets) for accurate analysis [11]. In addition, hosts used to source or relay spam often reside on types of computers not easily accessible to researchers.

[1] Strictly speaking, Code Red II itself has *not* been endemic since 2001. Rather, our data shows that the original has died off as programmed on October 1st of each year, but new variants are released including CodeRed.F[6] with the die-off date altered to give the worm extended life.

In fact, a significant percentage of open proxies used to send spam correlate with computers that have also transmitted viruses as email attachments [14]. This is likely due to the fact that recent malware such as SoBig or MyDoom create email proxies on the infected host.

4) Network (Mis)Configurations: Misconfigurations, default network settings, static software configurations and other oddities in network configuration settings have the potential of creating large, consistent *attractors* that, while a tremendous nuisance for the affected network administrators, can incidentally prove quite valuable for network measurement. In addition, because these are not a form of malware, they will often reflect a different class of host demographics.

One recently documented example comes from the configuration of several types of NetGear routers, which had hardcoded into them the address for an NTP server at the University of Wisconsin. This resulted in predictable, periodic traffic from in excess of a half million addresses [21]. Another example concerns a singular IP address in the UCSD Network Telescope, which receives a huge amount of spurious traffic due to its use as a preconfigured source address in a popular DDoS attack tool.

B. Example Events

We now sketch four specific network events that have provided opportunities for opportunistic measurement, with an emphasis on providing concrete examples of spurious network events and highlighting characteristics useful for measurement purposes. Later we discuss preliminary results drawn from some of these events.

1) Code Red II : The Code Red II worm [17] was released August 4th, 2001, and remains an endemic source of Internet "background radiation". One facet of the worm that provides for opportunistic measurement is its logic for selecting addresses to probe. Instead of selecting 32-bit addresses uniformly, Code Red II preferentially scans local subnets. Each infectee scans within the same /16 block as its own address with probability $\frac{3}{8}$; within the same /8 block with probability $\frac{1}{2}$; and within the entire Internet address space with probability $\frac{1}{8}$.

Another distinctive behavior of Code Red II is that it varies the number of threads it uses for scanning based on the *language setting* of the infectee's operating system, using 600 threads for systems with a setting of Chinese and 300 threads otherwise. Thus, the observed scan rate of a source could in principle be used to infer the language setting of the source, including use of non-Chinese systems inside of China, and vice versa.

2) Witty Worm: The Witty worm [18] was released in March, 2004. It spread worldwide, infecting 12,000 hosts, in 75 minutes. Witty was noteworthy in a number of ways: the entire worm fit within a single UDP packet; it was released the day after the announcement of the vulnerability it exploited; it was the first (and so far only) large-scale Internet worm that carried a destructive payload; and it targeted a flaw in the passive analysis of a network security product. Witty's basic structure was to seed the random number generator using the current uptime; generate and transmit 20,000 infection packets targeting randomly selected destination addresses; pick a random disk to corrupt; if the disk existed, corrupt a random block and start over with reseeding; otherwise, continue for another 20,000 infection packets without reseeding.

3) The Daily Eurasian Scan: A large-scale scan from sources in China (primarily), Korea, Japan, and Germany probes much of the IPv4 address space on a daily basis. Over time we have identified 16,000 sources participating in the scan, visible each day at CAIDA, Stanford, and LBNL. The scan sends SYN packets in bulk to ports 9898/tcp (Dabber Worm backdoor), 1023/tcp (Sasser Worm backdoor) and 5554/tcp (Sasser worm FTP server). Two characteristics of the scan make it exploitable for measurement: it predictably appears daily always within the same 5 minute time frame, and it generates large volumes of traffic. Because it consistently sends data at the same time each day, we can use it for studies sensitive to cross-traffic (such as queue occupancy and bandwidth estimations) without consideration of diurnal traffic variations. At the time of this writing it has been observed in traces for over a year [22].

4) A Heavily Spammed Domain: We acquired access to a long-standing Internet domain name that appears in widely distributed documentation. As a result, the domain receives up to 1 million spam emails daily, from over 40,000 source addresses. These spam emails can be large enough to produce TCP flows of sufficient size for complex flow analysis [11]. Furthermore, by changing the DNS records associated with the domain, we can *move* the focus of this traffic to different locations.

IV. Adapting Existing Measurement Methods

Current measurement practices typically either analyze traffic injected into the network in a controlled manner, or characterize network properties by investigating their direct effect on traffic. Opportunistic measurement, on the other hand, is "parasitic": the goal is to leverage existing traffic to infer unrelated properties of the network or the sending hosts. Even more than existing measurement techniques, however, opportunistic measurement is complicated by vagaries of the collection environment and limited knowledge of the sending hosts. In addition, it requires gaining access to useful vantage points for collecting traffic. In this section we discuss ways in which opportunistic measurement can allow us to more broadly apply existing analysis techniques.

A. Calibration

Unless we perform experiments in a highly controlled test-network, measurement studies must compensate for common disturbances, including drop rates, network outages, filtering rules, routing flaps, and queuing delays in the network. This is often done by having detailed knowledge and/or control of the sending source, and comparing sent traffic to received traffic.

With spurious traffic this is not practical, as the researcher has no control over the sources and likely little knowledge of network properties (such as filtering rules) between the traffic sources and the collection network. However, spurious traffic sometimes provides ways to calibrate measurements. Pulsars can exhibit quite reliable traffic patterns, with many topologically diverse sources and well-understood traffic coverage distributions. These may allow determination of whether biases exist in the collection network. For example, worms or botnets that scan uniformly across the address space can over time reveal the presence of filtering of certain sources, protocols or ports by the consistent absence of the corresponding traffic at the point of measurement.

B. Locality Biases and Attractors

Sources of spurious traffic often select their destinations in a non-uniform fashion. Such biases can considerably complicate some forms of direct analysis because the non-uniformity must be taken into account when extrapolating from the measurement's viewpoint to the broader Internet. However, such locality biases can also provide opportunities to infer otherwise hidden information (see § V-B).

When a locality bias is extreme, we can think of it as a form of traffic "attractor." If we can gain access to the attractor, we can leverage it for extensive measurement. For example, consider a public-address block that resides just above one of the private-address blocks. Any malware that *(i)* performs sequential scanning using its local address as a starting point (e.g., Blaster [8]), and *(ii)* runs on a host assigned to the private-address block will very rapidly probe the adjacent public-address block. Thus, the particular public-address block provides *magnified* visiblity into the malware's activity.

The magnification can be quite great: measurements we have conducted at such a block show a "background radiation" rate *1,000 times higher* than seen on other address blocks. In the extreme, the attractor receives *all* of the spurious traffic, such as the previously mentioned NetGear bug that flooded the Wisconsin NTP server [21].

Attractors needn't be IP addresses. For example, a domain name can serve as a focal point for spurious traffic, such as the spam target discussed previously. An advantage of these types of attractors is that we can *move* them to different points in the topology by modifying the mapping between the attractor's name (e.g., DNS record) and the corresponding IP address.

Finally, if we can compare observations from both within a preferred destination region and outside, then we can sometimes use the difference measured between the two to infer properties of the spurious traffic sources. Again, see § V-B for an example.

C. Inferring Network Properties

Traditional passive measurement techniques can be used with spurious traffic to infer network properties such as queuing delay via packet-pair variance, bandwidth and capacity measurements [13], [11], Internet distance studies via TTL analysis and packet loss estimation. For example, we used the MultiQ tool from the M&M tool suite [11] to estimate the bottleneck link bandwidth of 24,698 "significant" flows from 2,269 spam sources collected over a 24 hour period, finding clear spikes at popular bandwidths (modem speed, 10 Mbps Ethernet, 100 Mbps Ethernet). Particularly interesting was a clear, large spike at OC-12 speeds, which is consistent with the claim we have heard expressed that some spammers lease access to high-speed links to source their traffic.

D. Inferring End-Host and Edge Properties

Spurious traffic can provide detailed information about the end-hosts that source it. One particularly fruitful method for gathering end-host information is by exploiting the logic of the event source itself by reverse-engineering the code used to generate the traffic. See in particular the discussion of analyzing the Witty worm outbreak in § V-A below.

We can also use traditional passive techniques for inferring end-host characteristics. For example, we can infer the operating system and link speeds of the thousands of hosts present in a spurious traffic event using passive fingerprinting tools [3], [5] and packet-pair analysis [13], [11]. We can also sometimes exploit knowledge of the types of platforms associated with specific events for additional inference. For example, we might detect middleboxes by comparing the passive fingerprint of a source's traffic with the types of platform known to emanate the event type; in our study of Code Red II—a Windows-only worm—our analysis of traces using p0f [3] suggests that 14% of the 1,528 sources in fact originated from non-Windows sources, presumably reflecting Web proxies. We might also be able to detect and quantify middleboxes by the presence of their side effects, such as "VIA" headers in Web requests.

E. Amplifying Sources

While opportunistic measurement can be constrained by our level of access to spurious sources, it may be possible to coerce sources into sending traffic at higher rates (*agitating*) or attract new sources to a collection site (*chumming*).

Agitating a source of spurious traffic can cause it to send additional traffic, with a common example being responding to probes received at unused address spaces to increase the probability of enticing an incoming connection. For example, in one of our measurement environments we use high-interaction Windows honeypots that engage in full application conversations on TCP ports 135, 139 and 445, as well as UDP ports 137 and 138. Turning on the response mechanism increases the incoming packet count by over an order of magnitude for both TCP and UDP.

Chumming attempts to attract new spurious sources to a collection site. A simple example is intentionally disseminating an email address to mailing lists or placing it on the Web to increase its visibility to spammers. In addition, some network services are natural traffic attractors, such as the tendency of large IRC servers to become DDoS attack targets [19] due to their high visibility and popularity with script kiddies.

V. Preliminary Results

We consider two studies we performed recently using opportunistic measurement that demonstrate its utility and provide a taste of its use in practice.

A. Witty Worm Analysis

We begin with a recap of a recent analysis of the outbreak of the Witty worm. As reported in [12], with colleagues we analyzed traces from a /8 network telescope [15] that recorded approximately 1 in every 256 packets sent by each Witty infectee. Inspection of the worm's code (readily available since the worm sent a copy of itself in each packet!) revealed that it generated four 32-bit random numbers for each packet it sent: two from which it constructed a random destination address (using the top 16-bits of each random number), one for a random destination port address (Witty was unusual in that it triggered its exploit based on source port rather than destination port), and one for padding each packet to a varying degree.

However, the "random" numbers were of course not truly random, but rather pseudo-randomly generated using a linear congruential random number generator. The presence of bits from multiple random numbers in each packet then meant that it is possible to recover the state of the random number generator by inspecting a single packet sent by a Witty infectee; thus, the operation of the infectee ceases to be "random" and instead becomes completely deterministic and predictable.

Such determinism has great power. Because we can tell exactly how many packets an infectee sent between any two that appeared at the telescope, and because the system call used by infectees to send packets would block if the local link was busy, we can compute the bandwidth of the local link with high accuracy (as the volume of data sent between the two observed packets divided by the time elapsed between their observations), even in the presence of major packet loss between the source and the telescope.

Because Witty would reseed every 20,000 packets *only* if a randomly picked disk drive number corresponded to an actual disk, by observing the presence or absence of reseeding we can compute how many disks were attached to each infectee.

Finally, because Witty used the system uptime to see its random number generator, and because we could determine the seed by looking for linear (in time) increases in candidates present at reseeding events, we can also determine to high precision each infectee's uptime.

All three of these properties—access link bandwidth, number of attached disks, and uptime—are seemingly unmeasurable using any traditional techniques, yet we can measure them, and for a large population, using opportunistic measurement that exploits the exact structure of the source sending the spurious traffic.

B. NAT Usage Estimation using Code Red II Infectees

We have performed another study using opportunistic measurement to quantify the extent of NAT deployment for more than a thousand hosts infected by Code Red II. To do so we exploit the worm's preferential scanning of nearby prefixes. We tracked Code Red II scans over several monitored subnets and then computed the average number of probes received per /24. If a particular subnet receives, on average, approximately 2^{10} times more packets per /24 than others from a given source, then with high probability the source and the preferred subnet share the same /8.

The reasoning behind the factor of 2^{10} is as follows. As described earlier, $\frac{3}{8}$'s of the time Code Red II would generate a random address within its own /16; $\frac{1}{2}$ of the time, it would generate one within its own /8; and the remaining $\frac{1}{8}$, uniformly from the Internet's entire address space. Thus, $\frac{3}{8}$'s of the time, we have no opportunity to see the infectee's probe (unless the infectee was within the same /16 as one of our measured networks). The remaining $\frac{5}{8}$'s of the time, it would either target the same /8 resulting in $\frac{1}{2}$ packets going to 2^{16} /24s or the Internet at large with $\frac{1}{8}$ of the packets going to 2^{24} /24s. Therefore a /24 on a preferred subnet receives approximately $\frac{2^{27}}{2^{17}}$ times as many packets[2].

If the infectee had a public address, then we immediately know its address (because it's given as the source of the probes it sent), and can verify whether it's in the same block as one of our measured networks. If on the other hand the infectee had a private address, then we could *infer* this fact by its preference for /8 addresses near its private address.

For our study we concurrently collected a set of 48-hour HTTP traces. We monitored four /24 subnets and four /16 subnets that do not share a /8 with a private address block, six /24 subnets within the 192/8 space, and one /16 within the 169/8 space. To determine that a source was a Code Red II infectee, we searched the trace files for the specific HTTP GET request that Code Red II issues once it finds a host with port 80 open. From this we constructed a list of unique Code Red II source addresses and used it to reduce our trace file to only include traffic from each of the infected sources. The resulting data set consists of 487,291 scans from 1,528 sources.

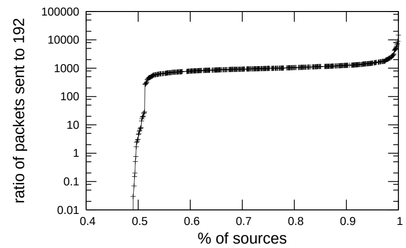

Fig. 1. CDF of the ratio of packets sent to /24s in the 192 /8 to /24s on other /8s.

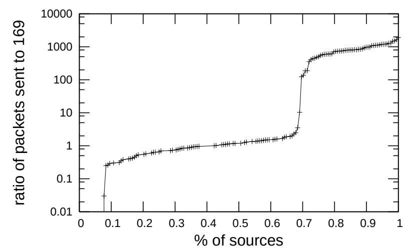

Fig. 2. CDF of the ratio of packets sent to /24s in the 169 /8 to /24s on other /8s.

Figure 1 shows a CDF of the ratio of probes from each source sent to the /24 blocks within the 192 /8 address space to /24s in other /8s. Based on our knowledge of Code Red II's scanning pattern, we expect an infectee within a 192.168/16 private address space to send approximately three orders of magnitude more of its probes to addresses within 192/8. Indeed, Figure 1 shows a clear mode of sources sending approximate 1000 times more of their scanning probes in this fashion. From the figure we see that around 50% of the infectees appear to use private IP addresses in the 192.168/16 range.[3]

We repeat the study with the sources that prefer the 192 subnet removed, looking for those that preferentially send traces to the /16 within the 169 subnet as shown in Figure 2. Roughly 70% of the remaining sources show no preference, however again there is a clear mode (approx. 30%) that prefer the 169 /8.

These results only provide a partial insight into private address usage by Code Red II infectees. Other reserved blocks such as the 10/8 and 172.16/12 are not represented. But, clearly, the infectees exhibit a great deal of private address use. We also need to remain cautious in generalizing these results as more broadly representative: Code Red II only infects hosts that have not been updated in several years. On the flip side, it is difficult for a host behind a NAT to become infected with Code Red II in the first place, since the NAT will not usually forward the incoming connection request.

[2] to be exact, a preferred /24 will receive slightly more than 2^{10} times as many packets due to packets it receives that are part of the $\frac{1}{8}$ sent to random addresses

[3] The irregularities in the figure concerning sources never sending or always sending to a 192/8 address arise due to granularity effects caused by the limited number of probes seen from many of the sources.

VI. LIMITATIONS

Many of the challenges of opportunistic measurement stem from lack of control over the traffic sources—a researcher only has passive access to pre-existing events. The researcher is thus constrained by the nature of traffic generated by existing events, the number of active sources, the periodicity during which the traffic is accessible, and the number of sources contributing. Furthermore, some events may only produce useful hotspots in areas of the Internet difficult to access (e.g., a commercially owned IP prefix).

We are further constrained by limited knowledge of the sending host. Unless inferable from the traffic, it is unlikely we can determine useful information regarding the sending host's operating system, hardware capabilities, connectivity bandwidth, upstream filtering policies, or intervening middleboxes.

Spurious traffic sources rarely provide an unbiased sampling of hosts on the Internet. One-time flash events, such as the outbreak of an Internet worm, typically emanate from a particular system type and configuration. Longer-lived worms, such as Code Red II, are more likely to remain on the machines of unsophisticated home users or small businesses. Machines that host bots, open mail proxies or other traffic source resulting from malware are more likely to remain active on machines with less administrative attention. A similar argument applies to end-host misconfigurations.

Dealing with these challenges remains a major hurdle for researchers to make effective use of opportunistic measurement.

VII. CONCLUSIONS

Advancement in empirical science is often tied to serendipitous reinterpretations of existing data—an accidental insight that transforms noise into meaning. Indeed, disciplines ranging from astronomy to medicine are rarely able to directly observe phenomena of interest. Instead, they must exploit secondary data sources to infer the hidden underlying activity—whether it be inflation via red shift or human infection via white cell count. In many ways these endeavors are apt metaphors for the challenges in Internet measurement as well. Few properties of the Internet can be directly measured themselves and thus the key innovations in the field are all inference-driven. Unfortunately, the Internet is changing in ways that make broad measurement increasingly untenable. Large swathes of the network cannot be probed directly and a variety of valuable measurement techniques have side-effects that are not socially permissible.

In this paper, we discuss a widely neglected data source with great potential to address these problems. Worms, scans, spam and DDoS are all scourges that routinely shower the Internet with spurious and unwanted traffic. While our ultimate hope is that such activity can eventually be eliminated, in the meantime we believe this traffic presents a viable and *unique* source of Internet measurement data. When one has lemons, one makes lemonade. Moreover, in our preliminary analyses we have found that exploiting such data allows a broader view (tens of thousands of hosts at a time), with greater diversity (in geography, topology, bandwidth and addressibility) and more useful traffic patterns (e.g., large TCP streams) than would be feasible with conventional network measurement approaches. While we acknowledge that these sources present their own difficulties and limitations, we believe it is exactly these challenges that our community is best at solving.

VIII. ACKNOWLEDGEMENTS

We would like to thank Colleen Shannon and Mark Allman for their helpful comments on this paper. Part of this research was performed while on appointment as a U.S. Department of Homeland Security (DHS) Fellow under the DHS Scholarship and Fellowship Program, a program administered by the Oak Ridge Institute for Science and Education (ORISE) for DHS through an interagency agreement with the U.S Department of Energy (DOE). ORISE is managed by Oak Ridge Associated Universities under DOE contract number DE-AC05-00OR22750. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of DHS, DOE, ORISE, or NSF. Support for this work was also provided by the National Science Foundation under grants CCF-0424422, CNS-0433668, CCR-0311690, NSF-0433702, and ITR/ANI-0205519, and by gifts from Microsoft Research, HP Labs, and VMWare.

REFERENCES

[1] CERT advisory CA-2001-26 nimda worm. http://www.cert.org/advisories/CA-2001-26.html.

[2] The dimes homepage. http://www.netdimes.org.

[3] p0f homepage. http://lcamtuf.coredump.cx/p0f.shtml.

[4] The planet-lab. http://www.planet-lab.org/.

[5] The siphon project: The passive network mapping tool. http://siphon.datanerds.net/.

[6] Codered.f. http://securityresponse.symantec.com/avcenter/venc/data/codered.f.html, March 2003.

[7] R. Caceres, N. Duffield, A. Feldmann, J. Friedmann, A. Greenberg, R. Greer, T. Johnson, C. Kalmanek, B. Krishnamurthy, D. Lavelle, P. Mishra, K. Ramakrishnan, J. Rexford, F. True, and J. van der Merwe. Measurement and analysis of ip network usage and behaviour. *IEEE Communications Magazine*, pages 144–151, May 2000.

[8] eEye Digital Security. Analysis:blaster worm. http://www.eeye.com/html/research/advisories/AL20030811.html.

[9] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot. Packet-level traffic measurements from the sprint IP backbone. *IEEE Network*, 2003.

[10] C. R. S. Jr. and G. F. Riley. Neti@home: A distributed approach to collecting end-to-end network performance measurements. In *PAM*, pages 168–174, 2004.

[11] S. Katti, D. Katabi, E. Kohler, and J. Strauss. M&M: A passive toolkit for measuring, correlating, and tracking path characteristics. Technical Report MIT-CSAIL-TR-945, MIT Computer Science and Artificial Intelligence Laboratory, April 2004.

[12] A. Kumar, V. Paxson, and N. Weaver. Exploiting underlying structure for detailed reconstruction of an internet-scale event. In *Proc. ACM IMC*, October 2005.

[13] K. Lai and M. Baker. Nettimer: A tool for measuring bottleneck link bandwidth. In *the USENIX Symposium on Internet Technologies and Systems.*, March 2001.

[14] R. Lemos. Spam may sprout viruses in home pcs. http://news.com.com/2100-1009-1021636.html.

[15] D. Moore. Network telescopes: Observing small or distant security events. Invited talk for USENIX Security, 2002.

[16] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver. Inside the slammer worm. *IEEE Security and Privacy*, 1(4):33–39, 2003.

[17] D. Moore and C. Shannon. The spread of the code-red worm (crv2). http://www.caida.org/analysis/security/code-red/coderedv2_analysis.xml.

[18] D. Moore and C. Shannon. The spread of the witty worm. http://www.caida.org/analysis/security/witty/.

[19] T. Olavsrud. Could attack on dalnet spell end for irc? http://www.internetnews.com/dev-news/article.php/10792_1573551.

[20] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of internet background radiation, October 2004.

[21] D. Plonka. Flawed routers flood university of wisconsin internet time server. http://www.cs.wisc.edu/~plonka/netgear-sntp/.

[22] C. Shannon. Personal correspondence, July 2005.