Project Summary: **A Real-time Lens into Dark Address Space of the Internet**

In the last decade, *network telescopes* have been used to observe unsolicited Internet traffic ("background radiation") sent to unassigned address space ("darkspace"). Network telescopes are one of the few types of instrumentation that allow global visibility into and historical trend analysis of a wide range of security-related events, including scanning address space for vulnerable targets, random spoofed source denial-of-service attacks, the automated spread of malicious software such as Internet worms or viruses, and miscellaneous misconfigurations. In recent years, traffic destined to darkspace has evolved to include longer-duration, low-intensity events intended to establish and maintain botnets. We propose to expand our telescope instrumentation to enable researchers to exploit this unique global data source to improve our understanding of security-related events such as large-scale attacks and malware spread.

Three pervasive challenges in network traffic research, including the telescope traffic, guide our proposed expansion: collection and storage, efficient curation, and sharing large volumes of data. The volume of data captured by the telescope is expensive to store, limiting the number of researchers who can realistically download data sets. The situation is worse during malicious activity outbreaks when the data volumes increase sharply, yet rapid analysis and response are necessary. Perhaps the most challenging obstacles to sharing any kind of Internet traffic data (even data to unused addresses!) are the privacy and security concerns. Viruses and worms may involve the installation of backdoors that provide unfettered access to infected computers, and telescope data could advertise these especially vulnerable machines.

We propose to deploy and evaluate an innovative shift in network monitoring that explicitly addresses all three challenges: enable near-real-time sharing of traffic data, in a way that maximizes data utility for research and analysis while protecting user privacy. We will improve classification of traffic to use a more modern taxonomy, including classes of DoS attacks, vulnerability scans, and malware spread. A meaningful taxonomy will help to create triggers to detect and notify interested researchers of events that merit more comprehensive measurement and analysis. We will also build infrastructure to allow vetted researchers to run analysis programs approximately one hour after data collection. For safe and ethical data sharing, we will use our recent Privacy-Sensitive Sharing Framework (PS2) which integrates privacy-enhancing technology with a policy framework using proven and standard privacy principles and obligations of data seekers and data providers.

The **intellectual merit** of our proposal lies in our proposed methodology and instrumentation enhancements that will increase the utility of network telescope instrumentation, transforming it into a more accessible, practically useful source of security-relevant data. The results of this project will contribute to developing efficient early detection, reaction and mitigation strategies thus enabling more scientific pursuit of cybersecurity research and critical advances in the global fight against pervasive malware.

The **broader impacts** of this project are diverse. We will broadly disseminate the results of this project to academic and security experts community via conferences, web sites, blogs and by organizing the proposed workshop. By creating educational data kits out of samples of telescope data containing security event signatures, this project creates an immediate link between research and education. Most importantly, it promises convenient remote access to a wealth of data, high-level computing resources and expertise of CAIDA researchers, lowering barriers to engaging in network security research for institutions serving underrepresented minorities.

Keywords: Network Measurement; Internet Darkspace Traffic Monitoring; Privacy-Sensitive Data Sharing

Project Description: II-EN: A Real-time Lens into Dark Address Space of the Internet

1 Motivation and Goals

In the last decade, *network telescopes* have been used to observe Internet "background radiation", i.e. unsolicited traffic sent to unassigned address space ("darkspace") [1]. The routing system carries the traffic to darkspace because its address is being announced globally, but there is no response back to the traffic sources since there are no hosts in darkspace. Observing such "one-way" traffic allows visibility into a wide range of security-related events, including scanning of address space by hackers looking for vulnerable targets, backscatter from denial-of-service attacks using random spoofed source addresses, the automated spread of malware such as worms or viruses, and various misconfigurations (e.g., mistyping an IP address). The observed packets represent mostly failed attempts to open connections, or other malware-related behavior. In the last two years this type of traffic has significantly increased due to botnet-related activities such as Conficker's scanning and p2p signaling [2, 3]. This year we have also seen a surprising increase in UDP traffic carrying payload that matches the signaling behavior of popular p2p file-sharing software [4]. We propose to expand our telescope instrumentation to enable researchers to exploit this unique global data source to improve our understanding of security-related events such as large-scale attacks and malware spread, and provide researchers and educators opportunities for real-time and historical analyses otherwise inaccessible, and at relatively little cost.

Three pervasive challenges in network traffic research, including on the telescope, guide our proposed expansion: collection and storage, efficient curation, and sharing large volumes of data. The volume of data captured by the telescope is expensive to store, limiting the number of researchers who can realistically download and process data sets. The situation is worse during malicious activity outbreaks when the data volumes increase sharply, yet rapid analysis and response are necessary[1].

Perhaps the most challenging obstacles to sharing any kind of Internet traffic data (even data to unused addresses!) are the privacy and security concerns. Viruses and worms may involve the installation of backdoors that provide unfettered access to infected computers, and telescope data could advertise these especially vulnerable machines. Yet ultimately, the speed, scope, and strength of today's automated malicious software demand effective real-time sources of data that can match the dynamics of the threat. Studying a worm in situ requires real-time traffic access, including raw victim host IP addresses and payload data.

We propose to deploy and evaluate an innovative shift in network monitoring that explicitly addresses all three challenges: implement real-time sharing of network traffic data, in a way that maximizes research and education utility while protecting user privacy. We will use a set of attributes we developed for real-time classification of traffic into known types (e.g., DoS attacks, vulnerability scans, etc.) to refine our methods and tools for early detection of meaningful changes in Internet background radiation. Improved reporting of statistics and event triggers and notifications will help researchers understand the macroscopic dynamics of the traffic and draw their attention to aspects of the traffic they may want to study while the event is still happening. To enable such responsiveness, we will enhance the infrastructure for telescope data collection and storage to allow vetted researchers to run analysis programs approximately one hour after data collection. We will support safe and ethical data sharing via our Privacy-Sensitive Sharing Framework (PS2) which integrates privacy-enhancing technology with a policy framework using

[1] Pang *et al.* described ways to filter traffic to reduce storage and processing costs, all of which come at some expense to potential research utility, see Section 2.

proven and standard privacy principles and obligations of data seekers and data providers. We will apply our PS2 model to the UCSD network telescope data and evaluate its applicability to other sources of data.

To our knowledge, such advanced model of network traffic data-sharing has never been tried before in the academic community, and there are logistic, policy, as well as technical challenges to overcome. But the outcomes we seek – more effective traffic monitoring instrumentation, and privacy frameworks to support sharing the resulting data – can help transform the relatively siloed, below-the-radar data sharing practices of the network and security research community into a more reputable and pervasive scientific discipline.

Section 2 describes our current data collection infrastructure and methodology for processing gigabytes of data from the UCSD Network Telescope and prior results. In Section 3, we propose to expand: (1) the technological capability of our darkspace (telescope) instrumentation, including a traffic classification and analysis methodology that will facilitate detection of changes in the nature of Internet background radiation over both long and short time scales; (2) hardware and software support real-time sharing of telescope data; (3) access and usability of gathered data through community development activities, including a new data-sharing policy framework that effectively manages privacy risks, and workshop for researchers interested in using our telescope data, improving telescopes as scientific instrumentation, or sharing their own telescope or other types of data in real-time.

2 Current status of data collection

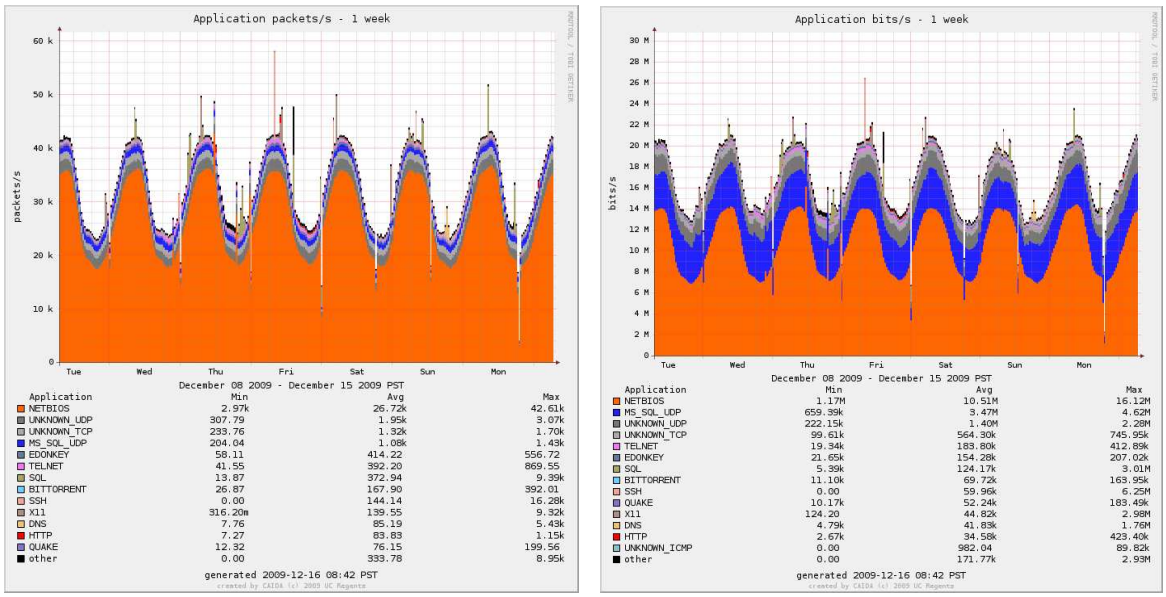
2.1 Existing telescope instrumentation

A telescope's *resolution* depends on the size of its address base; a telescope using a /16 address prefix will see more packets than one using a /24 prefix. The UCSD network telescope [5] uses a /8 mostly "dark" (unassigned) network prefix which corresponds to 1/256th of the total IPv4 address space and has only a few assigned addresses. We separate the legitimate traffic destined to those few reachable IP addresses, and monitor only the traffic destined to the empty address space. Therefore, if a host sends packets to uniformly random Internet (IPv4) addresses, the UCSD telescope should see about 1/256 of those probes.

We collect traffic data from the telescope using standard network interface cards in a PC, and CAIDA's Coralreef [6] software suite, which stores files in pcap format. As of December 2009, the network telescope captures in the range of 2GB up to and exceeding 100GB of compressed trace data per day. At the head-end of the capture process, the network's border router separates the legitimate traffic arriving at the telescope network (typically less than 1% of the total traffic volume) and forwards only non-legitimate traffic for monitoring and storage[2].

We have recently deployed a single host dedicated to storing approximately 30 days of data (at current traffic rates) from the UCSD Network Telescope. During normal operation the dataset covers a 30-most-recent-days moving window. Each day is represented by 24 compressed pcap files each containing one hour of data. Every hour the system automatically adds the most recent trace file to the collection, creating an almost real-time dataset with an effective latency of one hour. We periodically delete the oldest data to maintain the window without exhausting disk space. We hope to eventually increase our storage resources to maintain a larger window, since some event may not be recognized in its first 30 days of activity.

[2]The legitimate traffic is also a potential research resource, the sharing of which our PS2 framework [7] (cf. Section 3.3.1) could support.



(a) Packets vs. time (b) Bits vs. time

Figure 1: *Packets (left) and bits (right) to telescope for a week in December 2009. (Application is determined by sets of TCP/UDP port number, hence a very rough estimate.) The majority of traffic volume is TCP SYNs to port 445 (*NETBIOS*), which we stopped rate-limiting in April 2009 to observe Conficker.*

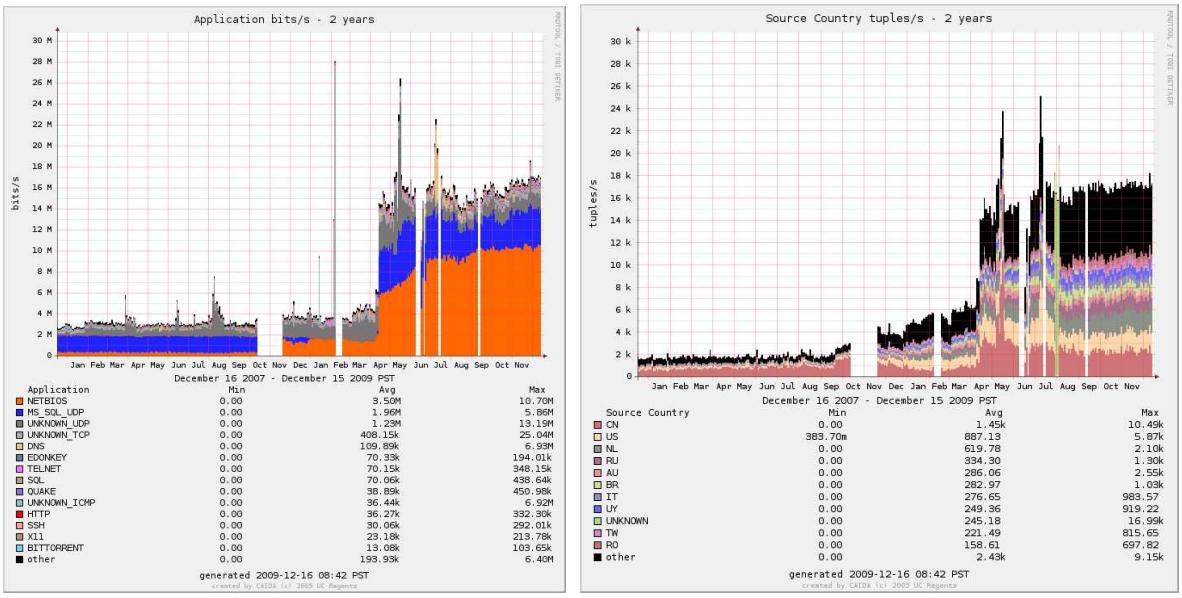
2.2 Reporting software

We provide a near-realtime, interactive graphical web interface [8] that displays rudimentary statistics of the telescope traffic. This coarse-grained view allows researchers to look for intervals of interest for further investigation.

Figures 1 (a,b) show the cross-section of packets and bits by application (determined by TCP/UDP port number, hence a very rough estimate) for a week in mid-December 2009. The vast majority of packets and bytes are TCP SYNs to port 445 (which strongly dominates the *NETBIOS* port category). However, larger UDP packets (in blue) constitute an increasing fraction of recent traffic, observable in Figure 2a, which plots traffic volume (bit) over the last two years. Figure 2a also reflects the fact that in April 2009 we removed a 2Mbps rate-limiting filter on packets sent to certain targeted ports (in place to reduce measurement load, a method also suggested in [1]), to obtain more accurate estimates of the Conficker spread. Rather than re-instantiate the filter, we are transitioning to a real-time sharing model that will allow researchers to create their own filters on traffic. The right plot on Figure 2b shows the tuples (5-tuple flows [9]) per second colored by source country (geolocated with NetAcuity [10]) over the last 2 years, showing an increase in probes from China and Russia, as well as from IP addresses we could not geolocate to a country.

2.3 Prior data releases

Over the years we released a number of general telescope datasets such as Backscatter data [11] as well as curated datasets focused on specific security events: Witty Worm public and restricted data [12], and Code-Red Worms data [13]. CAIDA has ameliorated the privacy risk of releasing victim host IP addresses and unexpected but occasional payload content with strict disclosure controls at some cost to research utility: (1) we deleted or sanitized payload; (2) we anonymized IP addresses of hosts using a common prefix-preserving technique. Our Acceptable Use Policies



(a) Traffic volume (b) Source country of IP source address

Figure 2: *Bits (left) per application, showing the removal of our rate-limit in April 2009, and the subsequent continued rise of packets probing previously rate-limited ports; Flows (right) per (geolocation-estimated with NetAcuity [10]) source countries sending to telescope for last 2 years, showing showing an increase in probes from China, Russia, and from IP addresses we could not geolocate to a country.*

control re-identification risk by requiring that researchers agree to make no attempts to reverse engineer, decrypt, or otherwise identify the original IP addresses collected in the trace [7].

Last year, we released two days of traffic traces from the network telescope in November 2008, as a baseline from dates prior to known Conficker activity, and then three days of data during Conficker growth [14, 15]. ***These datasets represent the only raw network telescope data that we know of for use by the academic research community, and a sample of the kind of data we propose to make available in more useful form and timing to researchers.***

2.4 Prior research and education contributions from telescope data

Network telescope data have allowed a few researchers – those with access to the data – visibility into a wide range of security-related events on the Internet, including misconfigurations malicious scanning of address space by hackers looking for vulnerable targets, backscatter from random source denial-of-service attacks, and the automated spread of malicious software called Internet worms [16, 17, 18, 19, 20, 21]. In a 2004 study, Pang *et al.* termed unsolicited packets to unassigned Internet addresses "*Internet Background Radiation* [1]", and analyzed a broad sample of such traffic to find that, relative to legitimate traffic, it "is complex in structure, highly automated, frequently malicious, potentially adversarial, and mutates at a rapid pace." [1]. Cooke *et al.* also described diversity in incoming traffic to ten unused address blocks [22] ranging in size from a /25 to a /8, announced from service provider networks, a large enterprise, and academic networks. Barford, *et al.* [23] found a bursty distribution of source addresses in darkspace, but consistency over time within a given chunk of dark space, and consistent (often vulnerable) destination ports probed, as expected for the targeted nature of most background radiation.

Prior research contributions enabled by data from UCSD's network telescope have included

studies of on DOS attacks [24, 25], Internet worms [26] and their victims, e.g., Code-Red [27], Slammer [28], and Witty [29] worms. Data sets curated from telescope observations of these events became a foundation for modeling the top speed of flash worms [30], the "worst-case scenario" economic damages from such a worm [31], and the pathways of their spread and potential means of defense [32]. CAIDA and other researchers continue studying the variety of traffic, and sources sending it, to empty address space.

CAIDA also participates in DHS's Protected REpository for the Defense of Infrastructure against Cyber Threats (PREDICT) project, intended to promote empirical research into network infrastructure security. Partial support is provided by PREDICT for annotating and indexing telescope data in the PREDICT meta-data repository, which will enable us to leverage NSF infrastructure funds and add synergy and cross-agency momentum to this proposed project.

3 Proposed Work

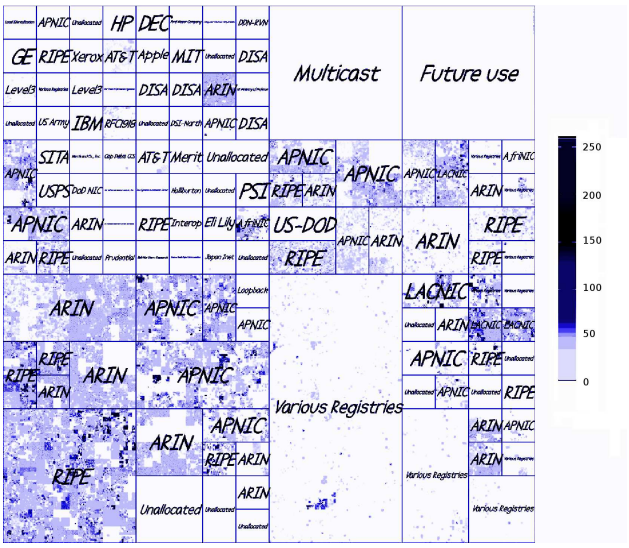
One of our most important lessons from Conficker is that in order to make the network telescope a more valuable resource for the next unpredictable large-scale outbreak of malicious network activity, we should be able to provide data access to vetted people with minimal curation and clean-up effort on our side, drastically reducing the lag time between data collection and analysis by researchers. In pursuit of this goal, we propose to enable experimental real-time data-sharing of telescope data. Our proposed tasks are: (1) enhance telescope data collection software and instrumentation to support automated generation of statistics, event triggers, and improved reporting focused discovery of emerging security threats; (2) deploy and evaluate a platform for real-time sharing of darkspace traffic data with vetted network and security researchers; (3) test a novel data-sharing framework and organize a workshop to get feedback from telescope data users.

3.1 Task 1: Enhance tools for telescope data analysis and visualization

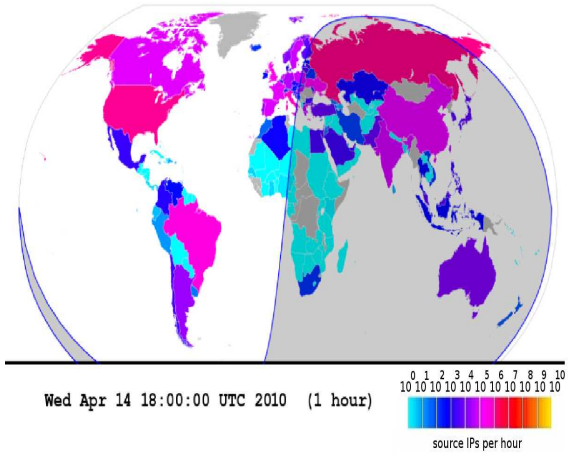
To increase the telescope usefulness as a way of detecting new security threats, we will extend our methodologies to support early detection and visualization of changes in the character of Internet background radiation. Specifically, we want to compute and report statistics at three layers: network, transport, and application. Network layer statistics include characterizing the geographic and topological origin of darkspace traffic. Transport layer (TCP and UDP header) information allows classification of behavioral patterns exhibited by individual IP source addresses. We have experimented with application-layer information extracted from the small fraction of darkspace traffic with payload to pattern-match against known set of application signatures, an area of stunted research due to the limited availability of traffic data. We expect that these tools will evolve based on research and experimentation enabled by our instrumentation enhancements.

3.1.1 Network-layer analysis

We propose to enhance the network-layer analysis operationally provided by the telescope, focusing on the topological and geographic distribution of traffic sources as indicated by the source address in the IP header. Figure 3a presents an example of aggregated day of data in a graphical *heatmap*, mapping IP source addresses to IANA IPv4 allocation blocks using Duane Wessels' heatmap plotting software [33]. The software maps the 1-dimensional IPv4 address space into a 2-dimensional image using a 12th-order Hilbert curve, so that CIDR network blocks appear as contiguous squares or rectangles in the image. Each point represents a single /24 network containing up to 256 hosts, and the color reflects observation of traffic from addresses in that /24.



(a) Heatmap of source IP addresses received on 14 Apr 2010. Each point represents a /24 network; color intensity reflects how many addresses in that /24 were observed sending traffic to the telescope. Most observed source addresses belong to blocks allocated by the RIRs; almost no addresses were coming from unallocated space, suggesting either a decrease in random spoofed source traffic or an increase in the sophistication of such spoofing.



(b) Geographic map of source IP addresses received on April 14, 2010. The color reflects how many addresses in that country were observed sending traffic to the telescope. There is approximate correlation to estimates of Internet users and usage in each country, suggesting that unsolicited traffic is truly a global phenomenon.

Figure 3: Hilbert heatmap and geographic heatmap of IP addresses transmitting to telescope.

These heatmaps [34] insightfully illustrate the non-uniformity of source addresses of traffic to the UCSD telescope; over time (e.g., animations of) such maps can reveal anomalous events. Most observed source addresses belong to blocks allocated by the Regional Internet Registries: RIPE, APNIC, ARIN, LACNIC, and AfriNIC; fewer are from areas labeled 'Various Registries', indicating allocations of pre-CIDR class B and class C IP address blocks before the mid-1990s. Even rarer are source addresses in the upper left of Figure 3a representing pre-CIDR class A blocks assigned to single organizations in the 1970s and 1980s. Almost no sources appear in Unallocated blocks, which suggests that (truly) randomly spoofed source addresses are an insignificant component of incoming traffic, consistent with previous work on darkspace source address distributions [23]. The heatmap technique [33] has become popular in the security community for visualizing penetration of a given vulnerability or attribute into the IPv4 address space. We will incorporate heatmaps into the daily telescope monitoring and reporting software.

Figure 3b maps the same data in geographic space on a world map.[3] The source addresses are well-distributed across the globe, with approximate correlation to estimates of Internet users and usage in each country, confirming that unsolicited traffic remains truly a global phenomenon. Geographic reporting is particularly important for identifying the origins of cybersecurity threats.

3.1.2 Transport-layer analysis

In search of computationally simple methods to distinguish among interesting classes of unsolicited traffic in close to real-time, we collaborated with U. Auckland computer science professor Nevil Brownlee and U Auckland's network security team to develop a taxonomy of *source types*.

[3] To map IP addresses to geographic locations we used Digital Envoy's NetAcuity service [10].

Group		Source Type		Description	% S	% p	kS	Mp
A	TCP	1	**TCP port probe**	TCP, many addrs, same port	14.66	53.08	242.59	64.67
		2	TCP only >1 port/addr	TCP, many addrs, >1 port	1.39	2.02	23.03	2.46
		3	TCP only 1 port/addr	TCP, 1 port for each addr	0.21	0.47	3.43	0.57
B	UDP	4	**UDP port probe**	UDP, many addrs, same port	27.31	1.74	451.84	2.13
		5	UDP only, >1 port/addr	UDP, many addrs, >1 port	7.02	7.7	116.11	9.39
		6	UDP only, 1 port/addr	UDP, 1 port for each addr	11.18	9.69	184.87	11.81
C	TCP+UDP	7	**Both TCP and UDP**	Mixed TCP and UDP	18.21	22.17	301.26	27.01
D	Conficker p2p	10	**Only Conficker p2p**	all pkts match Conficker p2p	3.53	1.30	58.39	1.58
		11	Mixed Conficker p2p	only some Conficker p2p pkts	0.17	0.26	2.78	0.32
E	Other	8	Other Protocols	no TCP or UDP (mainly ICMP)	0.06	0.38	1.07	0.47
		9	Backscatter	TCP with ACK flag set	0.01	0.11	0.17	0.14
F	Unclassified	0	unclassified	source sent less than 20 pkts	16.24	1.08	268.71	1.31

Table 1: Source taxonomy of one-way traffic as a function of transport protocol behavior, traffic volume, and address and port dispersion. The rightmost four columns show statistics on number and fractions of packets and unique source IP addresses observed at the UCSD Network Telescope in the hour from 00:00 on 1 Apr 2010 (UTC).

we experimented with many attribute combinations before settling on the manageable but meaningful taxonomy of source types (numbered 0 to 11) in Table 1, which we propose to to integrate into our operational real-time monitoring and reporting.

Because the number of sources in several of our *types* is low (<5%), we aggregated them into *source groups* based on transport-layer behavior attributes: transport protocol(s) used; number of unique destination addresses and ports; traffic volume; and consistency with known worm behavior. Our *source groups* are labeled A to F in Table 1. When all 20 packets from a source use the same transport protocol, TCP or UDP, we place that source into Group A or B, both of which generally correspond to probing of unassigned network addresses. Group C represents sources that generated both TCP and UDP packets in their first 20 packets, often reflecting applications that try to connect with TCP after failing with UDP connection attempts. Group E, 'Other', is mostly composed of ICMP packets and SYN ACK packets, which includes backscatter responses to spoofed addresses in our darkspace.

The four source types listed in bold in Table 1 show consistently distinct behavior, illustrated in Figure 4. Each vertical slice on these plots represents the distribution of that parameter for the corresponding hour, with color reflecting the number of sources matching each value on the y-axes. TCP Probe sources are either long-lived low-rate or short-lived high-rate; UDP Probe sources typically send about 10 packets in 5 seconds; TCP+UDP sources resemble the longer-lived TCP Probe sources; and Conficker p2p sources generated mostly long-lived, low intensity flows.

Conficker P2P sources illustrate the flexibility and power of our method in distinguishing among traffic behaviors, even for different behaviors from the same worm The bottom row of Figure 4 shows low-rate, long-lived sources searching for Conficker C p2p clients, only 17% of which had lifetimes <2800 s; most (61%) lasted >55 minutes. In contrast to Conficker A/B sources, which are captured by the TCP Probe source type with a daily peak of around 15:00 UTC (see also [35]), Conficker p2p sources have more complex behavior, with three peaks during workdays (suggesting peaks on different continents), but less apparent on weekends, suggesting that infected Conficker C hosts (or their networks) may be shut down on weekends.

We will further develop and refine this taxonomy, using it as the basis for both monitoring baseline trends in background radiation and developing triggers responsive to sudden changes.

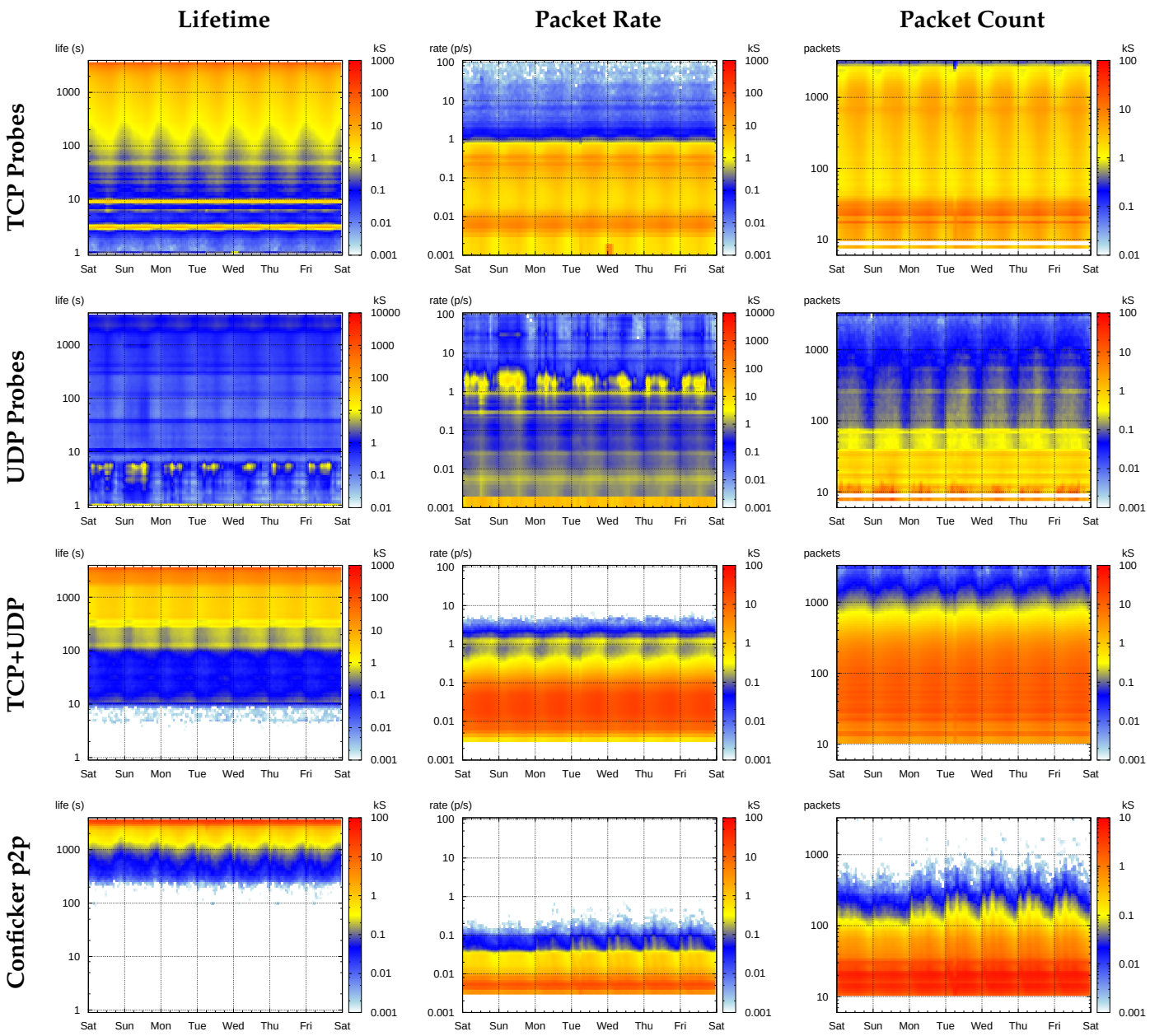


Figure 4: Hourly distributions of source lifetimes, packet rates and packet counts for four representative source types observed at the UCSD Network Telescope during the week Sat 3–Fri 9 April 2010 (UTC). Each type (1,4,7 and 10 in Table 1) shows notably *distinct* behavior patterns; although all have similar diurnal variations, their peaks appear at different values. Many sources are low-rate (below 0.01 p/s), but long-lived (lifetime peak near an hour), suggesting that they persist for periods greater than a single hour. TCP Probe sources are either long-lived low-rate or short-lived high-rate; UDP Probe sources typically send about 10 packets in 5 seconds; TCP+UDP sources resemble the longer-lived TCP Probe sources; and Conficker p2p sources generated mostly long-lived, low intensity flows.

3.1.3 Application-layer analysis

Given the privacy issues on host-populated networks, the telescope offers a unique opportunity to pursue payload-based traffic analysis, including evaluation and validation of traffic classification

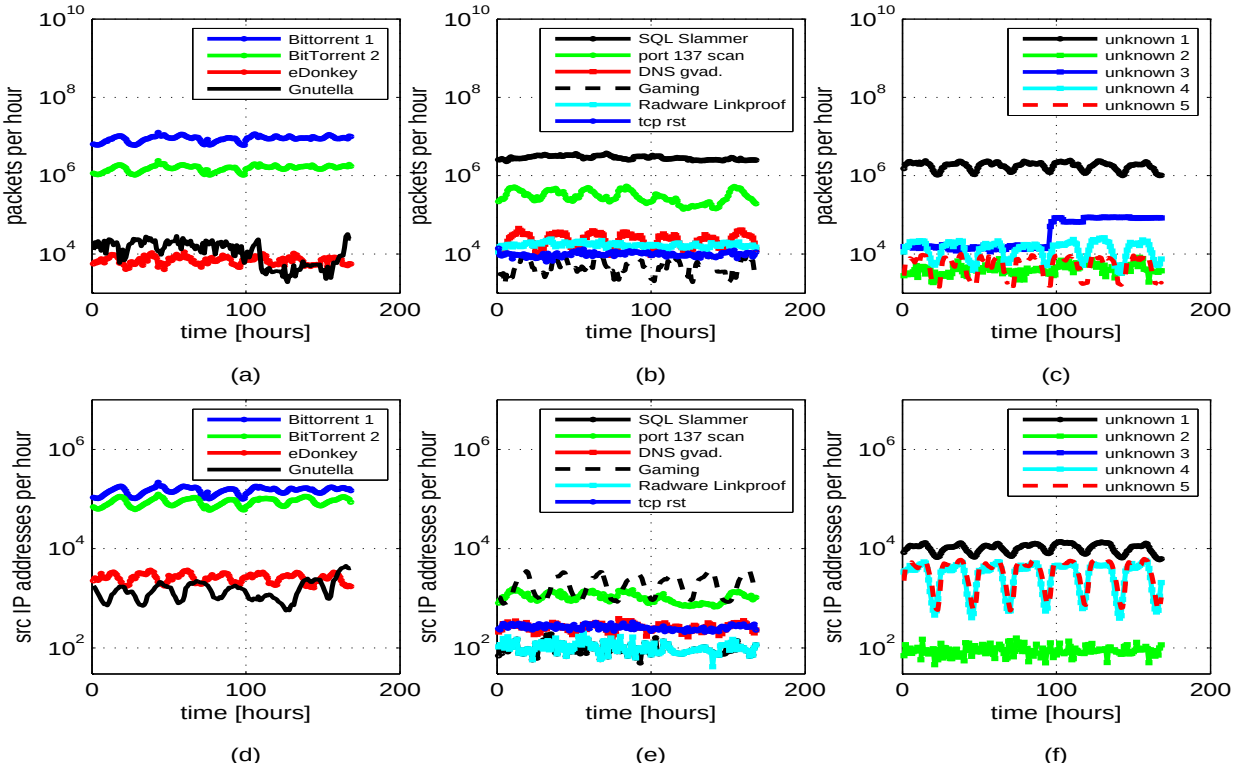


Figure 5: Packets containing payload, and IP addresses sending such packets, observed per hour at the telescope network during the week of 1-7 June 2010. Plots (a,d) show p2p protocols, plots (b,e) are for for known worms, scans, and other identified classes of malicious traffic, and plots (c,f) illustrate unknown classes requiring further study.

and modeling techniques. With respect for the sensitivities that remain in telescope data, we plan to carefully integrate application-layer analysis into our monitoring instrumentation.

To illustrate the possibilities, we analyzed a week of data from the telescope during 1-7 June 2010 to determine what security-relevant information can be extracted from payload carrying packets. More than 99.9% of the TCP packets (corresponding to the 78% of all received packets) in June 2010 carried no payload (mostly TCP SYNs), preventing application-layer analysis of those packets. However, almost all of the remaining 22% of (UDP) packets carried payload. To establish a baseline of traffic categories, we extracted transport protocol features (source and destination ports, payload size, and first 64 bytes of content) from a single day of data on 31 May 2010. We clustered all payload packets using a hierarchical clustering tool [36] to parameterize each class, then identified packets belonging to each class. Some classes matched protocol signatures in l7filter's set of known signatures or discoverable with web searches, but several classes we labeled "unknown".

Figure 5 shows the results of pattern-matching the June week of data against the classes extracted from the 31 May data. The top figure plots packets per hour matching each known class; the bottom plots unique source IPs sending those types of packets. Figures 5(a,d) show that a significant fraction of packets carrying payload bytes (more than 40%) matched patterns of p2p file-sharing protocols. The predominant packet pattern (indicated as "Bittorrent 1") was that used to initialize UDP BitTorrent connections, analogous to the TCP SYN three-way-handshake. Li *et al.* recently found that approximately 39% (in terms of the number of connections) of background

traffic that they monitored via honeynet sensors over the years 2004-2008 were the result of misconfigured p2p traffic [4], generated by software bugs and by injection of invalid peer addresses operated by anti-p2p companies attempting to pollute p2p networks. Figures 5(b,e) show packets matching patterns of known worms such as SQL Slammer on UDP port 1434 [28], the NetBIOS wildcard name query on UDP port 137 [37], repeated queries for the same DNS domain, and known software scan activities [38, 39]. We also include the rate of TCP RESET packets and corresponding IP sources addresses sending them.

This example demonstrates that payload analysis is another promising technique for characterizing the nature of background radiation. With careful attention to privacy issues, we will seek to integrate this technique with other types of monitoring that we propose.

3.2 Task 2: Deploy and refine hardware and software for real-time sharing of telescope data

Our previous model of indefinite storage and sharing of static, aged trace data on CAIDA servers [11, 12, 13] proved of some utility to cybersecurity researchers. Yet lessons learned from trying to share data during the Conficker attack motivate us to transition to a model of real-time data sharing with vetted researchers, storing a 30-day window (possibly up to 60 days since we recently increased our storage capabilities) of history. This new model should allow researchers timely access to a telescope observatory during a worm outbreak, where raw traces contain target addresses and payload that could enable quick autopsy of the structure and function of cybersecurity threats.

3.2.1 Hardware

We propose to harden the instrumentation by replacing the current NIC with a DAG card to prevent episodes of serious packet loss that cause inaccurate estimation of worm impact [40], as well as to increase timestamp precision. Both improvements would offer significant advantages to researchers analyzing the data.

To experiment with "bringing the code to the data" approaches, we will deploy a dedicated powerful data server capable to support several user accounts and their analysis tasks. We will also attempt to accommodate researchers who wish to provide their own computational resources, if our resources are insufficient for their needs. We propose to buy two fast network Juniper EX3200-48T-DC switches with two 10 gigE pluggable optics modules for each. These switches would replace our aging network hardware and raise our communication capacities to a new level thus making large volumes of data quickly and conveniently accessible for remote users and supporting additional equipment from other researchers.

3.2.2 User support

We will enable infrastructural and administrative support to open and maintain access for vetted researchers to the telescope data server. Our goal is to run analysis programs on data within approximately an hour of its collection. To support long-term historical analyses of trends and creation of educational data kits, we will archive periodic hourly and daily traces similar to our current telescope operations, with specific intervals based on community feedback. We will invite researchers and security experts working with similar and related data to help us evaluate our data collection parameters and curation methods. The PS2 framework described in Task 3 will provide guidelines on what information researchers may reveal vs. anonymize in analysis results.

We also anticipate that hands-on experience with real-time data analysis will prompt researchers to contribute to monitoring, analysis, and data visualization improvements to the telescope instru-

mentation (see also Section 3.3.2 on community outreach). We will evaluate such contributions to ensure that they meet appropriate ethics and safe data-sharing guidelines before integrating them into telescope operations.

3.3 Task 3: Community Development

3.3.1 Data-sharing policy framework

We have developed a Privacy-Sensitive Sharing Framework (PS2) [7] intended to effectively manage privacy risks that impede substantial data exchanges. The PS2 is a hybrid approach: a policy framework that applies proven and standard privacy principles between data seekers and data providers, coordinated with technologies that implement and enforce those obligations. The telescope data will be collected and made available as raw (unanonymized) traces, with payload (content). Rather than reduce the risk of releasing victim IPAs by anonymization and wholesale deletion of security-relevant data that also reduce the utility of the data to researchers, we will loosen the technical disclosure controls, and tighten use and disclosure obligations in the AUP.

We will enforce purpose specification by obtaining explicit web form acknowledgment from the researcher that s/he will use the dataset solely for the stated research purposes. Researchers also will have to consent to use appropriate and reasonable care in safeguarding access to and preventing unauthorized use of the data, and are explicitly not permitted to share or transfer their access to the data. They are also prohibited from attempting to connect to, probe, or in any other way interacting or intervening with a machine or machine administrator identified in the dataset, without permission from CAIDA. For any publication or other disclosure, the researcher is obligated to anonymize or aggregate IPAs, network names, and domain names unless obtaining written authorization from CAIDA to do otherwise. We will guard access to the dataset(s) and authorization to use it by application review, approval and communication of acquisition instructions by CAIDA administrators. We will also require the researcher to report to CAIDA summary of the research and any findings, publications, or URLs using the data.

3.3.2 Workshop on telescope measurements and data

In the beginning of Year 2 of the project we will host a workshop for researchers interested in using our telescope data, improving telescopes as scientific instrumentation, sharing their own telescope data, or sharing other types of data in real-time. This event, another in our series of popular Internet Statistics Measurements, and Analysis (ISMA) workshops [41], will help the community understand the breadth of research enabled by darkspace traffic data, as well as educate us regarding how much and what type of interest researchers have in darkspace traffic data, so we can maximize its utility for the research community. We will discuss lessons learned in this data-sharing experiment and publish a report summarizing findings and recommendations for use by other data providers and researchers.

4 Research and education opportunities enabled

Network telescopes are one of the only types of Internet measurement instrumentation that allows world-wide global view of Internet security-related phenomena, providing opportunities for real-time and historical analyses otherwise inaccessible, and at relatively little cost. The proposed infrastructure will enable novel studies of Internet security threats crucial for understanding their provenance and developing efficient mitigating strategies. Our enhancements allow for research

into network traffic in directions currently stunted for lack of access to such a data resource [42], including pursuit of the following questions:

1. What type of traffic is hitting the telescope *right now*?
2. Is aberrant traffic uniformly distributed across the address range?
3. Can we reliably taxonomize unsolicited traffic into specific categories (e.g., misconfiguration, backscatter, scan, worm, attacks), extract statistically significant trends, and catch emerging behavior patterns?
4. Which traffic characterization techniques will enable real-time anomaly detection?
5. How does this telescope darkspace compare to other darkspace, in traffic volumes, characteristics, stability, and patterns?
6. Can the proposed data-sharing model be used to allow such comparisons in real-time?
7. Is it possible to detect the onset of security-related events by real-time correlation of distributed information from a variety of other security measurement instrumentation, such as intrusion detection systems which face local threats and honeynets which directly engage with malware? Can we at least semi-automate the generation of local filtering policies, by inferring the local vs global nature of events?
8. How can we formally characterize the global nature and long-term trends in malware, e.g. worms, and the effectiveness of mitigation strategies, e.g. automated patching and malware removal? Do previous models of worm behavior match the unsolicited traffic observed by telescopes today, e.g., in terms of event duration and characteristics?
9. Can the telescope provide ongoing monitoring of the worldwide usage and trends in cyber-attacks? e.g., Are DOS attacks increasing in frequency and/or severity?
10. Which lessons of telescope data sharing can extend to other datasets, e.g. remote processing / real-time access?

The proposed instrumentation will also provide a fertile domain for **researchers, educators, and students to study spreading mechanisms, trends and evolution of malware on the global Internet**, while also exploring the policy aspects of sharing sensitive data sets. Our software tools will facilitate traffic analysis but also help annotate and anonymize traces creating specialized "educational data kits" focusing on signatures of specific malware.

The network telescope is an ideal place to test the PS2 framework, for many reasons: it is a privately owned network, operated for experimental uses, with no legitimate traffic collected, and we collaborate with the network owner on measurement for security purposes. We have also studied the extent of potentially personally identifiable information in telescope payload, and have not been able to identify any thus far. This project represents a unique opportunity to test, refine and further develop the PS2 framework based on practical experience. A data-sharing framework that effectively balances privacy risks and utility objectives promises far-reaching implications for an ever-increasing population of scientists working with privacy-sensitive data.

4.1 Why CAIDA is the most appropriate team for this project

CAIDA is a world leader in Internet measurement and data analysis of performance, workload, routing, topology, and security data [43], with years of experience in development, implementation, and evaluation of measurement infrastructure. We have been operating the UCSD telescope since 2001 and are committed to maintain and enhance this security research infrastructure. Support requested for this project combined with available DHS support for identifying, indexing, and annotating the most interesting subsets of data will ensure active life of the UCSD Network telescope until at least the end of 2013. beyond. SDSC provides access to advanced archival

storage systems, data-handling platforms, and high-bandwidth networking. This institutional infrastructure will extend the lifetime of procured data collection even longer, preserving them as a useful resource for the security research community.

5 Broader Impact Activities

5.1 Dissemination of results

The results of this project will be broadly disseminated and presented to both academic and operational security research communities. We will advertise the availability of telescope data via conferences, meetings, web pages and CAIDA's blog, and by organizing the proposed workshop. CAIDA is also an active participant in DHS's PREDICT project, which aims to publicize and provide datasets for cybersecurity research. We will also collaborate with security experts seeking to merge diverse datasets into a comprehensive multi-faceted characterization of existing and emerging Internet security threats, in support of the global fight against malware.

5.2 Integration of research and education

Some of the most valuable training for future researchers is not found in carefully controlled classroom experiences – real world data has unexpected problems. Telescope data kits will enable invaluable hands-on experience in operationally relevant network security and traffic analysis research. PI Kc Claffy will use an ongoing stream of current telescope data as educational aids for her teaching, both in and out of the classroom. She will mentor a graduate student to be hired for this project, teaching him/her the basic principles of Internet measurements and providing practical insights into the difficulties in working with of massive (and messy) real Internet datasets. Day-to-day experience with collecting, processing and documenting research data will expose the student to the range of problems that security datasets can have and instill in him/her a high level of scrutiny and healthy skepticism of unusual results.

CAIDA also has a successful track record of REU participation: in the last five years we supported and trained 17 undergraduates working on NSF-sponsored projects. We will apply for REU funds for this project as well.

Dr. Claffy will collaborate with professors from the UCSD Department of Computer Science and Engineering, Prof. Brownlee (University of Auckland, New Zealand), and with post-doc Alberto Dainotti of University of Napoli Federico II (and sometimes visiting scholar at UCSD/CAIDA) to develop hands-on class projects using anonymized telescope data kits.

5.3 Supporting Diversity with CAIDA Activities

Based at UC, San Diego, CAIDA has a strong record of integrating diversity into our research activities. Since 1999, the composition of our 90 paid interns has included 25 females, 21 Asians and 4 Hispanics. We attract a diversified pool of graduate students. In addition to advertising the available position on the CAIDA web site and communicating with our connections among faculty members of the UCSD Computer Science and Engineering Department, CAIDA will post through various specialized UCSD sites such as UCSD Society of Women Engineers (ucsdswe.org), UCSD Student Chapter of the Society of Hispanic Professional Engineers (shpe.ucsd.edu).

More importantly, resources like the UCSD Network Telescope and proposed real-time access mode are critical to the success of underrepresented groups in computer science and engineering, including women and minorities in other institutions. First, our telescope data are unique and the pool of unallocated IP addresses is rapidly shrinking, making it unfeasible to organize

a similar collection of blackhole space data elsewhere. Second, the necessary infrastructure for data collection, curation, and storage is extensive and expensive, difficult to set up and requires large initial investments. Finally, acquisition of data typically involves personal trust relationships and people-networking with engineering and management personnel, which can leave underrepresented groups at a social disadvantage. Yet our project offers easy universal remote access to available data sets, thus contributing a leveling influence to the research and education playing field and facilitating the entrance of women and minorities into network security research.

References

[1] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of Internet Background Radiation," in *Internet Measurement Conference (IMC)*, 2004.

[2] P. Porras, H. Saidi, and V. Yegneswaran, "Conficker," in *SRI International Technical Report*, 19 Mar 2000. http://mtc.sri.com/Conficker.

[3] P. Porras, H. Saidi, and V. Yegneswaran, "Conficker C P2P Protocol and Implementation," in *SRI International Technical Report*, 21 Sep 2009. http://mtc.sri.com/Conficker/P2P/.

[4] A. G. Z. Li, Y. Chen, and A. Kuzmanovic, "Measurement and Diagnosis of Address Misconfigured P2P Traffic," in *INFOCOM '10*, (San Diego, CA, USA), 2010.

[5] "UCSD Network Telescope," 2010. http://www.caida.org/data/passive/network_telescope.xml.

[6] D. Moore and K. Keys, "CoralReef software package." http://www.caida.org/tools/measurement/coralreef/.

[7] E. Kenneally and K. Claffy, "Dialing Privacy and Utility: A Proposed Data-sharing Framework to Advance Internet Research," *IEEE Security and Privacy (S&P)*, July 2010. http://www.caida.org/publications/papers/2009/dialing_privacy_utility/.

[8] "UCSD Network Telescope Monitor," 2010. http://www.caida.org/data/realtime/telescope/.

[9] K. Keys, D. Moore, R. Koga, E. Lagache, M. Tesch, and k. claffy, "The architecture of CoralReef: an Internet traffic monitoring software suite," in *PAM*, 2001. http://www.caida.org/outreach/papers/2001/CoralArch/.

[10] Digital Envoy, "Netacuity." http://www.digital-element.net/ip_intelligence/ip_intelligence.html.

[11] CAIDA, "Backscatter 2008 data." http://www.caida.org/data/passive/backscatter_2008_dataset.xml.

[12] CAIDA, "Witty data." http://www.caida.org/data/passive/witty_worm_dataset.xml.

[13] CAIDA, "Code red data." http://www.caida.org/data/passive/codered_worms_dataset.xml.

[14] "UCSD Network Telescope – Two Days in November 2008 Dataset," June 2009. http://www.caida.org/data/passive/telescope-2days-2008_dataset.xml.

[15] "UCSD Network Telescope – Three Days of Conficker Dataset," September 2009. http://www.caida.org/data/passive/telescope-3days-conficker_dataset.xml.

[16] M. Bailey, F. Jahanian, G. R. Malan, J. Nazario, D. Song, and R. Stone, "Measuring, Characterizing, and Tracking Internet Threat Dynamics," in *OpenSig 2003*, October.

[17] M. Bailey, E. Cooke, D. Watson, F. Jahanian, and J. Nazario, "The Blaster Worm: Then and Now," *IEEE Security & Privacy*, vol. 3, no. 4, pp. 26–31, 2005.

[18] D. Plonka, P. Barford, and V. Yegneswaran, "Internet Sink Deployments: On the Design and Use of Internet Sinks for Network Abuse Monitoring," in *IEPG Meeting*, July 2003. http://www.iepg.org/july2003/isink.pdf.

[19] T. Cymru, "Darknet project." http://www.team-cymru.org/Services/darknets.html.

[20] F. Jahanian, "Topology-Aware Internet Threat Detection Using Pervasive Darknets." http://www.eecs.umich.edu/fjgroup/topology/.

[21] A. Kumar, V. Paxson, and N. Weaver, "Exploiting underlying structure for detailed reconstruction of an Internet-scale event: Details of Witty worm," in *Internet Measurement Conference (IMC)*, pp. 351–364, 2005.

[22] E. Cooke, M. Bailey, Z. Mao, D. Watson, F. Jahanian, and D. McPherson, "Toward understanding distributed blackhole placement," in *Proc. ACM workshop on Rapid malcode, WORM '04*, (Washington DC, USA), pp. 54–64, 2004.

[23] P. Barford, R. Nowak, R. Willett, and V. Yegneswaran, "Toward a Model for Source Address of Internet Background Radiation," in *Proc. Passive and Active Measurement Conference, PAM '06*, (Adelaide, Australia), 2006.

[24] D. Moore, G. M. Voelker, and S. Savage, "Inferring Internet Denial-of-Service Activity," *Usenix Security Symposium*, 2001.

[25] D. Moore, C. Shannon, D. Brown, G. M. Voelker, and S. Savage, "Inferring Internet Denial-of-Service Activity," *ACM Transactions on Computer Systems*, 2004.

[26] Alberto Dainotti, Antonio Pescapé and G. Ventre, "Worm Traffic Analysis and Characterization," in *IEEE International Conference on Communications*, June 2007.

[27] D. Moore, C. Shannon, and J. Brown, "Code-Red: a case study on the spread and victims of an Internet worm," in *ACM Internet Measurement Workshop 2002*, Nov 2002.

[28] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the Slammer Worm," *IEEE Security and Privacy*, vol. 1, no. 4, pp. 33–39, 2003.

[29] D. Moore and C. Shannon, "The Spread of the Witty Worm," *IEEE Security and Privacy*, vol. 2, no. 4, pp. 46–50, 2005.

[30] S. Staniford, D. Moore, V. Paxson, and N. Weaver, "The top speed of flash worms," in *ACM Workshop on Rapid Malcode (WORM)*, pp. 33–42, 2004.

[31] N. Weaver and V. Paxson, "A worst-case worm," *Workshop on Economics and Information Security (WEIS)*, June 2004. `http://dtc.umn.edu/weis2004/weaver.pdf`.

[32] D. Moore, C. Shannon, G. Voelker, and S. Savage, "Internet Quarantine: Requirements for Containing Self-Propagating Code," in *INFOCOM03*, 2003. `http://www.caida.org/outreach/papers/2003/quarantine/`.

[33] D. Wessels, "IPv4-heatmaps." `http://maps.measurement-factory.com/`.

[34] A. Este, "IPv4 Address Space Maps of UCSD telescope data." `http://www.caida.org/~alice/docs/heatmaps_telescope.html`.

[35] E. Aben, "Conficker/Conflicker/Downadup as seen from UCSD Network Telescope, Feb 2009." `http://www.caida.org/research/security/ms08-067/conficker.xml`.

[36] M. Eisen and e. a. Michiel de Hoon, "Clustering 3.0." `http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm`.

[37] B. Alexander, "Port 137 scan." `http://www.sans.org/security-resources/idfaq/port_137.php`.

[38] E. Romang, "Activities on 49153/UDP linkproof.proximity.advanced ," April 2010. blog entry, `http://eromang.zataz.com/2010/04/28/suc007-activities-on-49153udp-linkproof-proximity-advanced/`.

[39] "Server queries." `http://www.gnu-darwin.org/www001/ports-1.5a-CURRENT/games/qstat/work/qstat-2.11/info/a2s.txt`.

[40] Songjie Wei and J. Mirkovic, "Correcting Congestion-Based Error in Network Telescopes Observations of Worm Dynamics," in *Internet Measurement Conference (IMC)*, 2008.

[41] "Internet Statistics and Metrics Analysis Workshops." `http://www.caida.org/outreach/workshops/isma/`.

[42] Alberto Dainotti, kc claffy, Antonio Pescapé, "Issues and Future Directions in Internet Traffic Classification." submitted to IEEE Network August 2010.

[43] CAIDA. `http://www.caida.org/outreach/papers/`.