Uses and Challenges for Network Datasets

John Heidemann
USC/ISI
johnh@isi.edu

Christos Papadopoulos
Colorado State University
christos@cs.colostate.edu

Abstract

Network datasets are necessary for many types of network research. While there has been significant discussion about specific datasets, there has been less about the overall state of network data collection. The goal of this paper is to explore the research questions facing the Internet today, the datasets needed to answer those questions, and the challenges to using those datasets. We suggest several practices that have proven important in use of current data sets, and open challenges to improve use of network data.

1 Introduction

Computer network research has long depended on a number of techniques, from reasoning and proof; to modeling and simulation; to experiments, from small-scale laboratories of a few PCs to large-scale testbeds such as Emulab [73] and PlanetLab [4]. While these tools all have a role, experience has shown that the Internet is inevitably more diverse and variable than we anticipate [55]. Direct study of the Internet itself is therefore an essential complement to the above tools—observation can provide the data to feed models, simulations, and experiments.

Unfortunately, direct observation of the Internet is quite challenging. The Internet is highly distributed—no central measurements have been possible since the NSFnet backbone was superseded in the mid-1990s [13]. In addition, as the Internet has been integrated with people's lives and businesses, very important privacy and legal protections have arisen [47].

Yet when gathered, data can be quite influential. As some examples: Paxon's study of pairwise TCP exchanges influenced TCP design and our understanding of network traffic [53]. Access to BGP routing updates made possible with the Route Views Project [58] and has advanced research and practice through scores of papers and new approaches to ISP relationships [20], routing [38], network efficiency [63], problem detection [12], and related topics. While relatively few packet traces are available, they have been very influential in denial-of-service, worm, and virus detection. While Route Views and packet traces shine light on only a fraction of the Internet today, their wide use and impact illustrates the promise of relevant network data.

The goal of this paper is to explore some of the research questions facing the Internet today, the datasets needed to answer those questions, and the challenges to using those datasets. Our thesis is that recently available datasets enable new research, but that continued work is needed to make new data available to address open research needs.

2 Research Questions

What are the key research questions that should drive Internet research today? A U.S. National Research Council report posed three grand goals: measuring the Internet, modeling networks, and disruptive prototypes [52]. Inspired by this report, CAIDA has conducted two "Day in the Life of the Internet" collection events [33]. But what are the more specific questions we should look for, and what data to answer them? Table 1 summarizes a number of important research topics, some of which we expand below.

Answering the overall question of what networking research is both very important and very difficult; we cannot possibly provide a complete description of the research space here. Instead, we seek to highlight a range of key questions that hopefully illustrate the space. We refer interested readers to the above NRC report [52] and other concurrent and subsequent reports for a fuller picture (for example, [60]).

2.1 Understanding Network Traffic

Network traffic has long been an area of study. Just as topology is studied at different levels, traffic has been studied as individual flows, aggregates on a link, or traffic matrices of the Internet.

Study of network traffic has had two main goals: characterizing typical traffic, and characterizing atypical, often malicious, traffic. The overall goal is to understand what

Table 1. General topics in network research and applications in those topics.

topic	applications
network traffic	
typical traffic	protocol design, congestion control, router buffer sizing, traffic modeling, new traffic types
atypical traffic	malware detection: denial-of-service attacks, worm, virus spread, malware; spyware; unusual traffic types; protocol verification
network topology	
AS-level	understanding business relationships
router- or link-level	evaluation of network robustness, cross-section throughput, network coordinate systems
address-level	evaluation of network size
topology and traffic	localizing attack sources, mapping network to geography, physical cross-section throughput

traffic dominates the Internet and how it affects traffic engineering, network architectures, and design of protocols, routers, firewalls, and other network appliances. Of course, understanding typical network traffic is a very broad research area; we provide a few representative examples below to illustrate how traces are used here.

Examples of studies of individual flows have often focused on TCP [53], or characterization of flows by size [69], duration [8], burstiness [59], or combinations of these [39]. Studies of individual flows are important to improve current protocols and understand how they will interact with new protocols. Studies of aggregate traffic include statistics of NSFnet [13], discovery of self-similarity in network traffic [40], to characterizations of traffic matrices [44]. An understanding of aggregate traffic is essential for medium- and long-term planning and traffic engineering [17, 19]. To date, study of network traffic has typically been done in consort with creation of new measurement infrastructure, or inside commercial ISPs. While the creation of trace infrastructure is understood relatively well today, we suggest that long-term evaluation of network trends requires analysis of through common datasets by multiple parties, something not generally possible today.

Internet traffic has been long studied; it might seem that there is little more to learn. However, the openness of the Internet means new applications constantly arise. Recent study of individual traffic has focused on new applications such as peer-to-peer file sharing [31, 14, 3], VOIP [6], YouTube video [10] and IPTV [11]. While studies of aggregate traffic have examined how these new applications have changed the traffic mix, we believe their is benefit to observing their behavior as they evolve from niche to mainstream. Part of the benefit includes new traffic models and user behavior, which are helpful in tuning anomaly detection.

Growth of malware and spam have prompted the detection of atypical Internet traffic. Several broad approaches have been considered, including entropy-based [22], change-point [64], parametric methods [24, 65] While these approaches use very different mathematical models to distinguish between typical and atypical traffic, *all* require examples of both attack and known non-attack traffic to evaluate their effectiveness and their rate of false positives.

Complementary to traffic detection, problems from malware can be avoided through protocol verification and software engineering techniques that reduce the number of bugs. In this case, traces can provide examples of protocols operating in unexpected ways, or of bugs being exploited.

In addition, characterizations of atypical traffic are constantly changing. Atypical traffic is often malicious; the adversarial relationship between malware authors and network operators leads to an arms race of continual evolution. This interaction stresses the importance of not just getting traces, but continuing to get new traces (Section 4.4).

2.2 Understanding Network Topology

That traffic must run over some network topology—a collection of links, LANs, and routers. Understanding the network topology is essential to understand the fragility or resilience of the Internet to attack or failure, developing models of network economics, developing services dependent on network location such as caching or replication, and other similar problems.

Internet topology was of wide interest in 2001–2002, with Skitter [27], Rocketfuel [62] and Mercator [63]. Studies considered both topology at the AS-level and router-level. Data from those studies has been essential to refine earlier observations about Internet topology [18] to reflect physical constraints [41].

Although the core Internet topology has been widely studied, some questions remain open, such as what is the least-cut diameter of the Internet, particularly when routing policies are considered. A full understanding of interactions between routing policy and raw connectivity is also pending.

We see a resurgence in interest in Internet topology, but now going to the edges of the network. While a number of groups have maintained manual lists of dynamically-assigned addresses, Xie et al. have inferred this information for accesses to e-mail provider logs [74]. Trestian et al. developed a classification method for addresses based on their presence on the web, as shown through the Google search engine [70]. We have been conducting census of all Internet addresses for several years [26]. Each of these approaches propose a new methodology and some direct applications; we believe their real power will arise as others apply the data in new ways.

Open questions in the core topology of the Internet are: how many Internet hosts are there, really? How many clients or servers?

2.3 Where Topology and Traffic Meet

Although traffic and topology have been studied in isolation, their combination provides a very compelling open area of research. The interaction of traffic is at the core of traffic engineering, and it also has bearing on policy issues such as network neutrality.

The traffic matrix is the first step in this direction [44], but traffic matrices have usually been studied only in the context of a single ISP. What is Internet-wide traffic like?

A second open area at the intersection of traffic and topology is to bring traffic into the physical world. Can we relate traffic with its geographic location? What will this tell us about caching policies, network provisioning, or geolocation of specific flows?

Finally, while the rest of Section 2 has focused on using network data to directly address problems facing the Internet today, there is an important *indirect* effect as well: data can be used to design, populate, and validate network simulations and models. Since simulation and modeling creates an isolated, malleable version of a subset Internet, it can be incredibly valuable in studying focused research questions. However, researchers must understand the relationship between what is modeled and the real world, if they expect their conclusions to reflect those constraints. The field of verification, validation, and accreditation of network simulations is an important area [25], yet one that can be challenging to apply in practice [37].

3 Classes of Data

Today, several research groups collect various types of data, both for their own research purposes and to provide data to the community through repositories such as PREDICT [68], for general Internet data, and CRAWDAD [35], for wireless networking. These systems store a large variety of data. As one example, PREDICT's privacy impact assessment lists 17 types of data [43].

To make sense of this data, we group these examples into several classes in Table 2 (this table is a new organization based on data assembled by Jody Westby and contributed by PREDICT participants). We consider data that is either local or network-wide, and either directly observed, or inferred from some analysis. Our focus in this paper and this table is on real-world datasets; we omit artificial or simulated approximations of these datasets here.

Each class of data can address different type of research questions. For example, locally observed data allows detailed drill-down into communication, providing a play-by-play account of security events, and in conjunction with packet-level traces, enabling the modeling or detection of malicious traffic. Locally observed events can provide a high-level description of ”what happened”, and network-wide observed data can observe global events such as worm outbreaks, routing failures and prefix hijacks.

The table lists providers of particular data types, both as part of the PREDICT program [68], or other public sources. It's important that several important data types are unavailable (to the best of our knowledge). Typically this data is limited because of privacy concerns; we expand on this point below in Section 4.2.

Finally, it is important that these data sources not simply be available in passing, but that there be datasets that large, representative, and public. Anyone can take full, unanonymized packet headers or system logs from their own computer, yet the generality of results drawn from such a dataset is much more limited than that taken from a large public network, or better still, from several different categories of large networks.

4 Lessons Learned

We next reflect on our experiences using trace data in research to find common problems that cut across types of data. We consider privacy issues, research requirements, and the nature of what we are observing.

4.1 Privacy and Anonymization

The data classes presented in Section 3 pose quite different privacy challenges; we next consider several categories of challenges.

First, observations that capture user data are most sensitive. This category includes locally observed data such as full packet contents (including data payloads). While potentially very useful, since it would enable deep packet inspection and so could provide ground truth for application or malware detection, this data is not currently available. The problem is that user data poses significant privacy and legal issues (see [47] for a discussion of these issues), thus such data can rarely be provided to researchers, and is almost

Table 2. List of data classes, instances of that class, and providers of that data (partially derived from data assembled by Jody Westby).

class	examples (formats)	Providers	Privacy Concerns
local observations	packet headers for general links (pcap or ERF)	CAIDA [9], LANDER [72], LBNL [54]	addresses
	packet headers for events such as attacks or worm spread (pcap or ERF)	CAIDA, LANDER, MERIT [71], LBNL	addresses
	full packet contents (pcap or ERF)	*unavailable*	user data and addresses
	flow-level traces (netflow)	MERIT	addresses
	router statistics (SNMP)	-	-
local inferences	intrusion detection alerts (Snort, Bro, etc.)	*unavailable*	addresses and system data
	logs (syslog, firewall, spam)	LogAnalysis.org [5]	addresses and system data
network-wide observations	active IP addresses	*unavailable*	general addresses
	DNS requests	*unavailable*	user data
	BGP tables	MERIT, RouteViews [58]	-
	end-host scans (ping or nmap)	LANDER	addresses
	topology scans (traceroute)	CAIDA	general addresses
	VOIP call records	PCH [48]	addresses
network-wide inferences	BGP hijackings (PHAS, bgpmon)	*unavailable*	-
	darknet/telescope packet headers (pcap)	CAIDA, MERIT	addresses
	darknet/telescope full packets (pcap)	CAIDA, MERIT	addresses and user data
	IP reputations	Spamhaus [1]	addresses

never shared without a legal warrant. We are not aware of any such data being available to researchers, even from their home institution.

Second is data that contains IP or MAC addresses (but not user data); examples include packet headers, flow records, and system logs. IP addresses pose a privacy concern because it is sometimes possible to relate them to the identities of individual humans; although not explicitly listed in relevant laws, there is general consensus that they constitute “personally identifying information” for purposes of U.S. and E.U. privacy laws. It is important to note that IP addresses *by themselves* do not identify users. Particularly with widespread use of dynamic addresses, IP addresses often must be combined with external information, such as user registration, DHCP logs, or application specific cookies, to map them to users. However, such mapping information is often maintained (sometimes to satisfy legal or operational requirements), and has been used under warrant to resolve IP addresses to users (even if incorrectly [56]).

Because of privacy concerns about IP addresses, several anonymization techniques have been proposed, such as prefix-preserving cryptographic-based renumbering [75]. Prefix preserving techniques are very useful to researchers because they preserve the structure of the network. Such approaches must be applied carefully, however, renumbering all user-specific fields (IP and MAC addresses) in headers and packet contents. Consistent renumbering schemes are also subject to attacks using external information (possibly injected by the attacker), or statistical analysis searching for well known, popular hosts [15], or common patterns such as sequential scans [49]. For MAC addresses, the options are to scramble the vendor and address portions of the addresses as one unit or independently [49].

Full address anonymization makes it difficult to associate traffic with organizations and makes some kinds of research impossible. For example, reverse engineering worm random-number generators [36] requires full, unanonymized addresses. An option is to anonymize or zero only part of the IP address (for example, as an optional in LANDER [28]). Such an approach confounds some number of addresses (256 or 65,000), balancing privacy while allowing traffic to be matched with large organizations. Matching data to organizations facilities some kinds of research, such as that comparing home, business, or academic use.

We classify two types of data (address scans and active IP addresses) as containing *general addresses*. By this term we mean they are IP addresses that are in use over some period but they are not associated with any other network information. In this sense, we believe they represent a reduced threat against privacy, particular if the period of use is broad (say, one week) and the survey size large. A good analogy is a list of phone numbers in a large city that were allocated or placed calls some time over a week—without specific times, call durations, or destinations, it is hard to see how they could be resolved to identification of an individual. Yet, there is significant value in this kind of information to addressing basic questions about Internet demographics and address utilization [26].

Finally, it is important to note that even without user data and with anonymized addresses, some information may leak. For example, OS fingerprinting tools such as p0f may still be used on a trace to determine the type of OS of a particular sender. Others have shown that clock drift can be analyzed and used to characterize hosts traces, including some based on clock drift analysis [34], or inference from regular patterns in scanners [49]. While researchers typically respond to fix explicit vulnerabilities relatively quickly, it is much harder to defend against unknown attacks that may be devised in the future [49].

The tension between analysis and privacy creates many difficult challenges. How much privacy are we willing to trade-off for better analysis? The answer can change drastically based on the context. This range of challenges has suggested that, rather than a simple policy (for example, renumber all IP addresses), a set of anonymization *rules* are required that can reflect context and more complex anonymization policies [50, 7, 49]. Such a framework is important, but leaves open how policies are to be defined. It also assumes well known, pre-defined data structures; such systems cannot cover transfer of unstructured or unknown user data, and so in these cases they must fall back to either removing, replacing with a hash, or encrypting such data. Finally, covert channels of information leakage may still exist even after anonymization. For example, identifying the busiest machine as a web or file server.

4.2 Unavailable Data

Table 2 lists several data types as currently unavailable. Typically this limitation is because of concerns about privacy, unknown approaches to anonymization, or availability of data only to commercial sources.

Full packet contents are unavailable because of clear privacy concerns. While it seems unlikely that full packet contents can be made available in general in any but unusual situations, some consideration has been given to how to anonymize full packet contents, although they have not yet reached their goal [49].

Local inference, either alerts from intrusion detection systems, or system logs is somewhat more promising. In fact, LogAnalysis provides sample system logs [5], but apparently with quite limited coverage. Other community supported repositories such as Dshield [66] provide user-contributed firewall logs and some analysis tools. Such efforts are invaluable, but information is not available in real-time making them better suited for post-mortem analysis. To the extent they represent criminal behavior (for example, spam, or break-in or denial-of-service attempts), it is possible they have lowered guarantees of privacy. However, potential false positives mean even this assumption must be taken carefully.

DNS requests, or other similar types of network infrastructure (perhaps NTP traffic) are again not currently available because of their uncertain privacy or anonymization methods. For example, DNS records suggest web browsing habits that could be tied to individuals or reveal sensitive information (for example, employees browsing job posting sites). Yet information about how network infrastructure is used is of great value in improving network operation. For example, early studies of DNS revealed several types of pathological behavior that significantly stress on the system [16], and recent work has used this infrastructure traffic to detect spam sources [30]. We hypothesize that they could be completely divorced of IP addresses and be left without any way to be identify individuals. However, caution is required, because prior experience suggests that often careful analysis can connect seemingly well-anonymized information to individuals, as shown when AOL's release of search engine records [23].

Finally, in most of these cases large datasets *do* exist, but they are only privately available. Often datasets are kept private because of legal concerns, concerns that they would release user or company private information, or because their owners consider them to represent commercially exploitable information.

4.3 Research Practices

We distinguish research from operations by its focus on developing new techniques as opposed to applying existing techniques to new networks. Since the goal of research is to discover something new, it poses two particular problems in data collection and analysis.

First, *validation* of new approaches is difficult yet essential. While it is relatively easy to guess and try new approaches, an approach is not solid science until researchers have been validated that it works, and more importantly, *why* it works. The process of defining a hypothesis, and testing it against known data to confirm it behaves as expected. While in some cases validation can be done "in the lab" with artificially generated data or simulations, comparisons against real-world data are often required to strengthen claims of correctness and accuracy (we give specific examples below). When comparing to ground truth derived from existing approaches or outside knowledge, *validation almost always requires less strict anonymization.* While future data collection may target strong anonymization for operations, it must leave opportunities for alternatives, at least with controlled subsets of data or populations of volunteers.

Two examples illustrate how validation often requires weaker anonymization. Spectral approaches to classification of denial-of-service attacks into single-source and distributed are advantageous because they are blind—the operate only on packet timing, not contents [29]. Developing this approach required knowledge of the behavior of single- and multi-source attacks, knowledge we could only get if they were already identified. In our case, we bootstrapped our analysis with header-based (non-blind) identification, and followed with simulation and experimental studies. But header-based analysis depends on more information (weaker anonymization) than bind spectral analysis—this approach to validation would have been impossible with a more adversarial foe. As a second approach, we developed a blind technique to detect peer-to-peer file sharing [3]. Ideally we would validate this against known peer-to-peer users, but strong evidence of such activity requires deep packet inspection. In this example, only packet headers were available to us, so we fell back on port-based identification, although we know many peer-to-peer approaches now avoid well known ports. This case illustrates how greater information would have lead to a more definitive result.

Second, *development of new approaches always requires iteration in what data is collected.* Unless one can store all observations for all time, data collection always requires summarizing what is observed, omitting some information. Anonymization adds another level of intentional omission. Unfortunately, all too often, important details of the data are omitted. Although researchers plan data summarization and anonymization carefully, research is, by definition, discovering the unknown, so these plans nearly always fall short. Chances of incomplete can be minimized by extensive planning before data collection, but this approach greatly increases the cost of research and inevitably decreases the ability to pursue interesting but unexpected phenomena. Finally, often researchers simply do not know what data is important until after several iterations (we thank an anonymous reviewer for this observation). We suggest that flexibility and correctness are best when researchers iterate with data collection and analysis, since iteration means that collection strategies can change as becomes necessary.

We found iteration was essential in our study of Internet address space usage [26]. Table 3 shows the evolution of the data we save, with significant changes as we tried to use the data to reach conclusions and found that what we saved was insufficient. We have been studying the Internet address space for five years, but our earliest measurements preserved very little information—just a bitmap of responsive addresses. While responsiveness is the most important information, it is far from the only information; we have extended our storage format four times to date. We added recording of negative (error) replies to understand measurement errors, and then later found negative replies reveal information about use of router access control lists. This result was an opportunistic side-effect of the core research made possible by iteration of analysis and data collection. Our current format is much more careful to save all data we receive, even unexpected or invalid data, for future evaluation. While it is possible we were overly naïve in our initial data formats, we think it is more likely that this kind of iteration is *inherent* to the development of data collection. Just as extensive use of software is part of debugging, use of data is essential to debugging what is collected.

4.4 A Moving Target and Coverage

Finally, we suggest that continued observation is important even when some data already exists.

First, *evolving areas of the Internet need continuous data collection* Data collection should not be considered a one-time activity, but needs to evolve as the Internet does. Repeated data collection is essential because most interesting aspects of the Internet continue to change. Malware, such as denial-of-service attacks and spam, provide a clear example of this problem. At one time DoS and spam were quite simple, depending on floods from a single host or using open mail relays. As defenses have improved, these attacks have evolved. As a result, traces showing this previously common behavior no longer reflect techniques currently in use.

Second, *multiple datasets of the same type provide additional value*. Because of the incredible diversity of the Internet [55], apparent redundancy in datasets provides an important ability to confirm observations from one dataset apply elsewhere. Furthermore, observations from any one location may be biased by the local traffic mix, network connectivity, or other factors. Multiple view points are essential. One specific example of this need has been seen in studies of AS topologies [57].

5 Open Research Directions

We have made significant progress in distributed datasets, but further work remains.

Better *anonymization* approaches are still needed. Although current prefix-preserving IP addresses anonymization seems to work reasonably well, provided care is taken, additional work is needed to understand how to anonymize other types of data, potentially including user data [49], or application-level headers. The ability of HTTP to pass through firewalls has made it a convenient encoding for non-web applications (including streaming media, RealAudio; voice-over-IP, Skype; virtual-private network protocols, and other media). There would be significant value to separating these uses of HTTP as a transport layer from

Table 3. Evolution of information saved in address scans.

version	year	information
0	2003	bit per responding addresses, for ICMP echo reply only
0.1	2004	adds TTL, RTT
1	2005	new format: encoded ICMP type and reply code (not all saved), TTL, RTT, for three ICMP message types only
2	2007	new format: full ICMP type and reply code, TTL, RTT, for all valid ICMP message types
2.1	2008	adds pcap capture of all invalid ICMP message types

HTTP as a web application. Better anonymization approaches have just begun to be explored in recent workshops [2, 51].

Complementing anonymization must be understanding of *privacy attacks*. As we describe above (Section 4.1), even widely used anonymization schemes leak some information. Understanding how to characterize and mitigate these kinds of attacks is essential, particularly if we are to explore weaker forms of anonymization for subsets of data. The database community has been successful establishing principles to contain information leakage (for example, [45, 42]). Networking researchers have just begun exploring how these approaches apply to trace analysis, and if new tools can contain information leakage [46].

Dataset *annotations and metadata* become increasingly important as datasets are used and researchers identify positive or negative features. This problem is well known in data curation; network-central systems such as DatCat provide facilities for shared annotations [61]. Metadata is particularly challenging because, as with what basic data to capture (Section 4.3), there are many details that could be captured and only the iteration of multiple research users identify what is important.

We have focused on datasets for the traditional, wired Internet. Data specific to *other access types*—wireless mesh networks, telephone networks, or even SCADA or sensor networks. Wireless and telephone networks are increasingly IP-based, but the different mix of applications and use patterns may influence observations. Some dataset providers have already focused on wireless-specific datasets [35].

Finally, although outside the scope of technical challenges, *revisiting the social and legal* scope of network tracing is important. Well understood best practices are needed in passive data observation and in active probing and participation: what is it acceptable to observe with what level of aggregation or anonymization? Even careful active probing incurs cost on the target, particularly when it is or could be misunderstood as malicious. What are standards for how to balance these costs and benefits? When is participation in a network of malware for research appropriate? When should network monitoring be subject to human-subjects review processes such as Institutional Review Boards [67, 21]? Finally, what are the legal frameworks for data collection, and what grey areas need clarification? And given that the Internet spans international borders, how does one consolidate legal frameworks from different countries? Early exploration here has begun to explore legal questions, but opened many more [47, 32].

6 Conclusion

This paper has outlined classes of available network datasets and how that data can support network research. While there is more data available today than in the past, supporting new kinds of research will require both new datasets and new approaches to managing anonymization, privacy, and the social framework of research. The key to moving research progress forward is the iteration between application-driven researcher needs (Section 2 and new approaches (Section 5) in the context of growing experience (Section 4). Finally, an important non-technical issue is developing the appropriate national and international legal framework for distributing network traces, necessitating close collaboration between researchers, lawyers and policy makers.

Acknowledgments

This work is partially supported by the U.S. Dept. of Homeland Security contract NBCHC080035 (LANDER-2007). This work is partially supported by the National Science Foundation (NSF) under grant number CNS-0823774, “NeTS-NBD: Maltraffic Analysis and Detection in Challenging and Aggregate Traffic (MADCAT)”. Conclusions of this work are those of the authors and do not necessarily reflect the views of DHS or NSF.

We thank Yuri Pradkin for assistance with Table 3 and this research, and Jody Westby and the members of the PREDICT program for their contributions to Table 2.

Finally, we thank the anonymous reviewers for their suggestions which significantly improved our paper.

References

[1] The Spamhaus project. web site `http://www.spamhaus.org/`.

[2] S. Antonatos, M. Bezzi, E. Boschi, B. Trammell, and W. Yurcik, editors. *Proceedings of the 1st ACM CCS Workshop on Network Data Anonymization*, Alexandria, VA, USA, Oct. 2008.

[3] G. Bartlett, J. Heidemann, and C. Papadopoulos. Inherent behaviors for on-line detection of peer-to-peer file sharing. In *Proceedings of the 10th IEEE Global Internet Symposium*, pages 55–60, Anchorage, Alaska, USA, May 2007. IEEE. An extended version of this paper is available as ISI-TR-2006-627.

[4] A. Bavier, N. Feamster, M. Huang, L. Peterson, and J. Rexford. In VINI veritas: Realistic and controlled network experimentation. In *Proceedings of the ACM SIGCOMM Conference*, pages 3–14, Pisa, Italy, Sept. 2006. ACM.

[5] T. Bird and M. J. Ranum. LogAnalysis.org. Web site `www.loganalysis.org`, 2002.

[6] R. Birke, M. Mellia, M. Petracca, and D. Rossi. Understanding VoIP from backbone measurements. In *Proceedings of the IEEE Infocom*, pages 2027–2035. IEEE, May 2007.

[7] M. Bishop, B. Bhumiratana, R. Crawford, and K. Levitt. How to sanitize data. In *Proceedings of the 13th IEEE Internation Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE)*, pages 217–222. IEEE, June 2004.

[8] N. Brownlee and kc claffy. Understanding internet traffic streams: Dragonflies and tortoises. submitted for publication, July 2002.

[9] CAIDA. Supporting research and development of security technologies through network and security data collection. Project described at `http://www.caida.org/funding/predict/`, Dec. 2007.

[10] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the ACM Internet Measurement Conference*, San Diego, CA, USA, Oct. 2007. ACM.

[11] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain. Watching television over an IP network. In *Proceedings of the ACM Internet Measurement Conference*, pages 71–84, Vouliagmeni, Greece, Oct. 2008. ACM.

[12] D.-F. Chang, R. Govindan, and J. Heidemann. Locating BGP missing routes using multiple perspectives. Technical Report ISI-TR-2004-588, USC/Information Sciences Institute, May 2004.

[13] K. Claffy, H.-W. Braun, and G. Polyzos. Traffic characteristics of the T1 NSFNET backbone. In *Proceedings of the IEEE Infocom*. IEEE, Jan. 1993. Also available as SDSC Report GA-A21019 UCSD Report CS92-237.

[14] F. Constantinou and P. Mavrommatis. Identifying known and unknown peer-to-peer traffic. In *Proceedings of the Fifth IEEE International Symposium on Network Computing and Applications*, pages 93–102, Cambridge, MA, USA, July 2006. IEEE.

[15] S. E. Coull, C. V. Wright, F. Monrose, M. P. Collins, and M. K. Reiter. Playing devil's advocate: Inferring sensitive information from anonymized network traces. In *Proceedings of the ISOC Network and Distributed System Security Symposium*, pages 41–48, Alexandria, Virgina, USA, Feb. 2007. The Internet Society.

[16] P. B. Danzig, K. Obraczka, and A. Kumar. An analysis of wide-area name server traffic: A study of the domain name system. In *Proceedings of the ACM SIGCOMM Conference*, pages 281–292, Jan. 1992.

[17] N. Duffield and M. Grossglauser. Trajectory sampling for direct traffic observation. In *Proceedings of the ACM SIGCOMM Conference*, pages 179–191, Stockholm, Sweeden, Aug. 2000. ACM.

[18] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the ACM SIGCOMM Conference*, pages 251–262, Cambridge, MA, USA, Sept. 1999. ACM.

[19] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving traffic demands for operational IP networks: Methodology and experience. In *Proceedings of the ACM SIGCOMM Conference*, pages 257–270, Stockholm, Sweeden, Aug. 2000. ACM.

[20] L. Gao. On inferring automonous system relationships in the internet. *ACM/IEEE Transactions on Networking*, 9(6):733–745, Dec. 2001.

[21] S. L. Garfinkel. IRBs and security research: Myths, facts and mission creep. In *Proceedings of the USENIX Usability, Psychology, and Security (UPSEC)*, San Francisco, CA, USA, Apr. 2008. USENIX.

[22] Y. Gu, A. McCallum, and D. Towsley. Detecting anomalies in network traffic using maximum entropy estimation. In *Proceedings of the ACM Internet Measurement Conference*, pages 345–350, Berkeley, California, USA, Oct. 2005. USENIX.

[23] K. Hafner. Researchers yearn to use AOL logs, but they hesitate. New York Times, 23 Aug. 2006, Aug. 2006.

[24] X. He, C. Papadopoulos, J. Heidemann, U. Mitra, and U. Riaz. Remote detection of bottleneck links using spectral and statistical methods. to appear, Computer Networks, Sept. 2008.

[25] J. Heidemann, K. Mills, and S. Kumar. Expanding confidence in network simulation. *IEEE Network Magazine*, 15(5):58–63, Sept./Oct. 2001.

[26] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and survey of the visible Internet. In *Proceedings of the ACM Internet Measurement Conference*, pages 169–182, Vouliagmeni, Greece, Oct. 2008. ACM.

[27] B. Huffaker, M. Fomenkov, D. Moore, and kc claffy. Macroscopic analyses of the infrastructure: measurement and visualization of internet connectivity and performance. `http://www.caida.org/outreach/papers/pam2001/skitter.xml`, Nov. 2001.

[28] A. Hussain, G. Bartlett, Y. Pryadkin, J. Heidemann, C. Papadopoulos, and J. Bannister. Experiences with a continuous network tracing infrastructure. In *Proceedings of the ACM SIGCOMM MineNet Workshop*, pages 185–190, Philadelphia, PA, USA, Aug. 2005. ACM.

[29] A. Hussain, J. Heidemann, and C. Papadopoulos. A framework for classifying denial of service attacks. In *Proceedings of the ACM SIGCOMM Conference*, pages 99–110, Karlsruhe, Germany, Aug. 2003. ACM.

[30] K. Ishibashi, T. Toyono, K. Toyama, M. Ishino, H. Ohshima, and I. Mizukoshi. Detecting mass-mailing worm infected

hosts by mining DNS traffic data. In *Proceedings of the ACM SIGCOMM MineNet Workshop*, pages 159–164, Philadelphia, PA, USA, Aug. 2005. ACM.

[31] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel traffic classification in the dark. In *Proceedings of the ACM SIGCOMM Conference*, pages 229–240, Philadelphia, PA, USA, Aug. 2005. ACM.

[32] kc claffy. Ten things lawyers should know about Internet research. Posted to the CAIDA blog, `http://blog.caida.org/best_available_data/2008/04/16/top-ten-things-law%yers-should-know-about-internet-research-1/`, Apr. 2008.

[33] kim claffy. Following up on 'A Day in the Life of the Internet' challenge. Blog entry `http://blog.caida.org/best_available_data/2007/06/20/following-up-a-day%-in-the-life/`, June 2007.

[34] T. Kohno, A. Broido, and K. C. Claffy. The devil and packet trace anonymization. In *Proceedings of the 2005 IEEE Symposium on Security and Privacy*, 2005.

[35] D. Kotz, T. Henderson, and I. Abyzov. CRAWDAD: A community resource for archiving wireless data at Dartmouth. web site `http://crawdad.cs.dartmouth.edu/`, Dec. 2004.

[36] A. Kumar, V. Paxson, and N. Weaver. Exploiting underlying structure for detailed reconstruction of an Internet-scale event. In *Proceedings of the ACM Internet Measurement Conference*, pages 33–47, Berkeley, California, USA, Oct. 2005. ACM.

[37] S. Kurkowski, T. Camp, and M. Colagrosso. MANET simulation studies: The incredibles. *ACM Mobile Computing and Communications Review*, 1(2):50–61, Oct. 2005.

[38] C. Labovitz, A. Ahuja, A. Abose, and F. Jahanian. Delayed Internet routing convergence. In *Proceedings of the ACM SIGCOMM Conference*, pages 175–187, Stockholm, Sweeden, Aug. 2000. ACM.

[39] K.-C. Lan and J. Heidemann. A measurement study of correlation of Internet flow characteristics. *Computer Networks*, 50(1):46–62, Jan. 2006.

[40] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *ACM/IEEE Transactions on Networking*, 2(1):1–15, Feb. 1994.

[41] L. Li, D. Alderson, W. Willinger, and J. Doyle. A first-principles approach to understanding the Internet's router-level topology. In *Proceedings of the ACM SIGCOMM Conference*, pages 3–14, Portland, Oregon, USA, Aug. 2004. ACM.

[42] A. Machanavajjhala and J. Gehrke. On the efficiency of checking perfect privacy. In *Proceedings of the 25th ACM Symposium on Principles of Database Systems*, pages 163–172, Chicago, Illinois, USA, June 2006. ACM.

[43] D. Maughan. Privacy impact assessment for the protected repository for the defense of infrastructure against cyber threats (PREDICT). DHS release, Feb. 2008.

[44] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic matrix estimation: Existing techniques and new directions. In *Proceedings of the ACM SIGCOMM Conference*, pages 161–174, Pittsburgh, Pennsylvania, USA, Oct. 2002. ACM.

[45] G. Miklau and D. Suciu. A formal analysis of information disclosure in data exchange. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 575–586, Paris, France, 2004. ACM.

[46] J. Mirkovic. Privacy-safe network trace sharing via secure queries. In *Proceedings of the ACM CCS Workshop on Network Data Anonymization*, pages 3–10, Alexandria, VA, USA, Oct. 2008. ACM.

[47] P. Ohm, D. Sicker, and D. Grunwald. Legal issues surrounding monitoring during network research (invited paper). In *Proceedings of the ACM Internet Measurement Conference*, pages 152–160, San Diego, CA, USA, Oct. 2007. ACM.

[48] Packet Clearing House. Packet clearing house website. `http://www.pch.net`, 2007.

[49] R. Pang, M. Allman, V. Paxson, and J. Lee. The devil and packet trace anonymization. *ACM Computer Communication Review*, 36(1):29–38, Jan. 2006.

[50] R. Pang and V. Paxson. A high-level programming environment for packet trace anonymization and transformation. In *ACM SIGCOMM Conference*, pages 339–351, Karlsruhe, Germany, Aug. 2003. ACM.

[51] C. Papadopoulos, editor. *Proceedings of the 22nd IEEE Conference on Computer Workstations*, Steamboat Springs, CO USA, Oct. 2008.

[52] D. Patterson (chair), editor. *Looking Over the Fence at Networks: A Neighbor's View of Networking Research*. National Academy Press, 2001.

[53] V. Paxson. End-to-end Internet packet dynamics. *ACM/IEEE Transactions on Networking*, 7(3):277–292, June 1999.

[54] V. Paxson. LBNL/ICSI enterprise tracing project. web page `http://www.icir.org/enterprise-tracing/`, 2005.

[55] V. Paxson and S. Floyd. Why we don't know how to simulate the Internet. In *Proceedings of the 29th SCS Winter Simulation Conference*, pages 1037–1044, Atlanta, Georga, USA, Dec. 1997. Society for Computer Simulation.

[56] M. Piatek, T. Kohno, and A. Krishnamurthy. Challenges and directions for monitoring P2P file sharing networks, or why my printer received a DMCA takedown notice. Technical Report 08-6-01, University of Washington Dept. of CSE, June 2008.

[57] M. Roughan, J. Tuke, and O. Maennel. Bigfoot, Sasquatch, the Yeti and other missing links: what we don't know about the AS graph. In *Proceedings of the ACM Internet Measurement Conference*, pages 325–330, Vouliagmeni, Greece, Oct. 2008. ACM.

[58] Route Views. University of Oregon Route Views Project. web site `http://www.routeviews.org`, 2000.

[59] S. Sarvotham, R. Riedi, and R. Baraniuk. Connection-level analysis and modelling of network traffic. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*, pages 99–103, San Francisco Bay Area, USA, Nov. 2001. ACM.

[60] E. Schmidt (chair), editor. *The Internet's Coming of Age*. National Academy Press, 2001.

[61] C. Shannon, D. Moore, K. Keys, M. Fomenkov, B. Huffaker, and kim claffy. The Internet measurement data catalog. *ACM Computer Communication Review*, 35(5):97–100, Oct. 2005.

[62] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with Rocketfuel. In *Proceedings of the ACM SIGCOMM Conference*, pages 133–145, Pittsburgh, Pennsylvania, USA, Aug. 2002. ACM.

[63] H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin. The impact of routing policy on Internet paths. In *Proceedings of the IEEE Infocom*, pages 736–742, Anchorage, Alaska, USA, Apr. 2001. IEEE.

[64] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blažek, and H. Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9):3372–3382, Sept. 2006.

[65] G. Thatte, U. Mitra, and J. Heidemann. Detection of low-rate attacks in computer networks. In *Proceedings of the 11th IEEE Global Internet Symposium*, pages 1–6, Phoenix, Arizona, USA, Apr. 2008. IEEE.

[66] The Internet Storm Center. Dshield: Cooperative network security community. web site `http://www.dsheld.org`, 2001.

[67] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Technical report, Department of Health, Education, and Welfare, 1979.

[68] The PREDICT Program. PREDICT portal overview. web page `https://www.predict.org/Portals/0/files/Documentation/MANUAL%20OF%20OPE%RATIONS/PREDICT_Overview_final.pdf`, Aug. 2006.

[69] K. Thompson, G. J. Miller, and R. Wilder. Wide-area internet traffic patterns and characteristics (extended version). *IEEE Network Magazine*, 11(6):10–23, Nov/Dec 1997.

[70] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Unconstrained endpoint profiling (Googling the Internet). In *Proceedings of the ACM SIGCOMM Conference*, pages 279–290, Seattle, Washigton, USA, Aug. 2008. ACM.

[71] University of Michigan. The virtual center for network and security data. web page `http://www.eecs.umich.edu/fjgroup/predict/`, 2007.

[72] USC/LANDER Project. LANDER: Los Angeles network data exchange and repository. web page `http://www.isi.edu/ant/lander/`, Nov. 2007.

[73] B. White, J. Lepreau, L. Stoller, R. Ricci, S. Guruprasad, M. Newbold, M. Hibler, C. Barb, and A. Joglekar. An integrated experimental environment for distributed systems and networks. In *Proceedings of the Fifth USENIX Symposium on Operating Systems Design and Implementation*, pages 255–270, Boston, Mass., USA, Dec. 2002. USENIX.

[74] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber. How dynamic are IP addresses? In *Proceedings of the ACM SIGCOMM Conference*, pages 301–312, Kyoto, Japan, Aug. 2007. ACM.

[75] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon. Prefix-preserving IP address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. In *Proceedings of the 10th IEEE International Conference on Network Protocols*, pages 280–289, Washington, DC, USA, Nov. 2002. IEEE.