Addressing Ethical Considerations in Network Measurement Papers

Craig Partridge
Raytheon BBN Technologies
craig@aland.bbn.com

Mark Allman
ICSI
mallman@icir.org

ABSTRACT

Network measurement—because it is typically at arm's length from human beings—does not comfortably fit into the usual human-centered models for evaluating ethical research practices. Nonetheless, the network measurement community is increasingly facing ethics issues and finding itself poorly prepared. We discuss why the ethical issues appear somewhat different for network measurement, and propose that measurement papers be required to include an ethical considerations section. We believe that some of the ideas here will also prove applicable to other areas of computing systems measurement, where the researcher's attempt to measure a system may have an impact on human beings.

1. INTRODUCTION

A conference program committee (PC) is usually the first outside independent organization to evaluate research work that measures network systems. In recent years, questions about whether the work submitted follows sound ethical practices have become increasingly common within PC discussions. The authors have experience with this situation as researchers, as well as members and leaders of PCs struggling with ethical concerns.

The fundamental cause of this struggle is that our community does not have a set of shared ethical norms. Historically, measurements of computing and communications systems have not been viewed as impacting humans to a degree that required ethical review. Indeed, ethics review boards often declare our research as "exempt" from full review as not involving human subjects ([5] describes a recent experience). Hence, outside the need to protect the privacy of communications content, we lack consensus about how to ethically handle even the most basic situations. Often authors work from one set of ethical notions while the PC applies one or more different sets of ethical underpinnings as part of their review. This leaves well meaning community members—in all roles—on fundamentally different pages. The situation is further exacerbated because our community does not have a culture of describing the ethical reasoning behind a set of experiments. This situation (i) leaves PCs to try to derive the foundations on which the paper stands and (ii) means that precautions taken by a careful researcher are not exposed to others who may leverage or build upon previous techniques in subsequent work.

In this article we advocate requiring an "ethical considerations" section in measurement papers as a first step in addressing these issues. By requiring such a section—even if the result is a statement that there are no ethical issues—we provide the starting point for a discussion about ethics in which (i) authors have a chance to justify the ethical foundations of their experimental methodologies and (ii) PC members can review the authors' perspective and can provide specific feedback as necessary. Further, by including these sections in published papers the entire community starts to develop a collective understanding of both what is ethically acceptable and how to think through ethics issues.[1]

Our aim in this article is to present an initial strawman. We do not attempt to prescribe what is and what is not ethical. We do not tackle all possible ethical questions that arise in our work as Internet empiricalists. Rather, we advocate for a framework to help the community start an explicit conversation about the largest ethical issues involved in measuring networked systems such as the Internet, and also cloud computing systems and distributed transactions systems.

2. BACKGROUND

There are three strands of intellectual activity that come together when one examines ethics and network measurement.

1. The evolving field of ethics in information and communication.

2. The evolution of the field of network measurement. Developing an empirical understanding of network behavior has been a pillar of network research from its earliest days. This research area has steadily improved, refining its tools to extract ever more information from measurements—such that longtime assumptions about what information can be extracted from a measurement often no longer holds.

3. The legal issues surrounding network measurement—a topic still in its infancy [16]. Legal issues are at best murky in a single jurisdiction since there is little case law to lay the foundation for how courts will interpret communication systems law within the context of modern data networks. Such issues multiply when a measurement study crosses (many) jurisdictions. We encourage researchers to consult their local counsel when legal questions arise. However, for the purposes of this article our focus is on ethical issues and we will mention legal issues only when they help illuminate the ethical issues.

[1] The important issue of exposing ethical issues raised by rejected papers remains. We are setting this issue aside in an effort to make some progress.

2.1 Ethics

The study of ethics in information and communication science has, broadly, followed two (overlapping) lines of inquiry.

The first line is a focus on human-centered values such as life, health, security and happiness. This thinking originated in the 1940s with Norbert Weiner and has carried down to the present. A current expression is the *Menlo Report*, a 2012 report by the US Department of Homeland Security [8]. The Menlo Report focuses on issues of causing harm to persons, either through revealing confidential information or altering their environment, and ensuring the risks of harm from the experiment are recognized, moderated and equitably distributed (e.g., seeking to ensure that those persons whose environment is altered by the experiment are also persons who are likely to benefit from the experimental results).

The other line of ethical thinking has focused on the professional responsibility of the computing and information sciences professional. Specifically the focus has been on following good industry practices in the creation of software artifacts, and codes of conduct that outline a professional's responsibilities to society, employer, colleagues and self. A detailed expression of this thinking is the joint IEEE/ACM *Software Engineering Code of Ethics and Professional Practice*, which identifies eight principles and presents over 80 distinct admonitions [2].

Both these approaches are concerned with the impact of one's work on other humans, and systems that directly interface with humans.

Network measurements, and many other system measurement, are usually at least one step removed from directly interfacing with humans. Intuitively, probing a network or counting hits in a cache does not impact humans. Nonetheless, measurement work can impact humans, and in this article we will focus on measurements where the human impact, however indirect, can clearly be envisaged. This focus means we will not focus on ethical issues where the harm, to first order, might come to vendors, systems, or intellectual property rights owners.

2.2 Evolution of Network Measurement

The field of network measurement—broadly defined—is relatively old. As best as we can tell, beginning with the electronic telegraph all networks have been the subject of various forms of measurement. We briefly trace the evolution of the field both technically and in a legal and ethical context and finish with some observations.

2.2.1 Technical Evolution of Measurement

Measuring a communications network and analyzing the results has been a staple of (data) communications research from its inception. By 1911, AT&T had a statistical group that, among other functions, leveraged measurement to better engineer the telephone system and to predict demand. When the ARPANET (forerunner of the Internet) was turned on in 1969, its first node was installed at UCLA for the use of Leonard Kleinrock's measurement group.

Measurement can be passive or active. Passive measurement simply observes in-situ traffic. Active measurement injects new traffic to observe the system's response. Given that networks are digital systems, built according to standards, a reader might imagine that examining network traffic (passively or actively) is largely an exercise in detecting bugs. In reality, the interactions of traffic in networks give rise to complex patterns. Furthermore, because the communications infrastructure is distributed, the interaction of delays between components and routine failures can lead to complex phenomena. Finally, variations in how specifications are implemented can lead to interesting interactions.

Examples of important research results from passive monitoring include: methods for ensuring sequence numbers are robust against device crashes [18], the discovery of self-similarity in network traffic [12], and methods to avoid the self-synchronization of network traffic [9]. Examples from active probing include measurements to develop the Network Time Protocol (which keeps distributed clocks synchronized) [13] and the study of network topology [20].

2.2.2 Ethics and Law of Measurement

Much of our legal, social and ethical dialog about network measurement uses legal terminology that was developed in the early days of measurement. Specifically, the ethics and legality of network measurements are often evaluated with the implicit assumption that the only parties who can capture data outside a workplace campus are communications companies providing service and Government agencies given access to communications companies' data centers (see for instance [1]). Further, a typical formulation distinguishes between two classes of data, as follows.

The first class of data reveals when and how long two parties communicated. United States law defines a device capable of capturing such data as a *pen register*. More recently, the term *metadata* has been used to describe an expanded set of information, including packet headers, that it is argued is comparable to pen register data.

The second class of data reveals the *contents* of the conversation. To highlight the distinction, consider a phone call to a bank. A pen register records that a call took place at a specific time and for a specific duration. The contents of the conversation would reveal that the call was, for example, a balance enquiry. United States law, since 1967, has recognized that the contents of a conversation are a distinct class of information that has a higher expectation of privacy [19], and this distinction between contents and metadata is often carried over into ethical discussions.

2.2.3 Metadata is Becoming Content

A variety of factors have eroded the distinction between content and metadata. Specifically, our ability to leverage metadata to infer—or even re-create—content is increasing rapidly.

A few examples illustrate this point:

- Measuring when devices in a network transmit is sufficient to derive traffic tables that delineate routers from end systems and identify which nodes are communicating with each other [6].

- *The Queue Inference Engine* takes information about transactions (e.g., pen register style data) and reverse engineers it to determine the behavior of queues [11]. Researchers have made steady progress in using techniques such as QIE to characterize queues from metadata. For instance, we can tell whether and roughly how long a person likely waited in line at a bank ATM machine by tracking when transactions at the machine start and end [3].

- Inter-packet gaps (metadata) between encrypted transmissions can be used to infer where users' fingers were on the keyboard and thus give guidance about what letters are in their passwords [17].

- In some cases, it is possible to determine what words are being said in an encrypted voice conversation, simply by looking at packet headers [21]

Summarizing, with less data than a pen register would collect, we can often determine that the call to the bank was a balance enquiry. Furthermore, we should expect the distinction between metadata to data to continue to erode over time.

3. THE CONTOURS OF HARM

While there are myriad ethical issues that confront network measurement work, our aim in this article is to address those causing *tangible harm to people*. We are not concerned with notions of potential harm to network resources (e.g., bandwidth) or equipment, except insofar as the impact on resources and equipment causes tangible harm to a human. We believe how our work impacts individual human beings is the most important ethical issue.

Additionally, we note that our goal—which agrees with the Menlo Report—is not to eliminate the possibility of harm within our experiments. Rather, we aim to minimize the risk of inflicting harm. In this context we make several observations which bear on how we manage risk in our experiments:

A Spectrum of Harm: First, we recognize that "harm" is difficult to define. Rather than a precise definition we offer that a single probe packet sent to an IP address constitutes at best *slight harm*.[2] Meanwhile, a persistent high-rate series of probes to a given IP address may well be viewed as both an attack and create *serious harm* (e.g., by clogging a link precisely when it is needed for an emergency). These ends of the spectrum are useful as touchstones when thinking about how to cope with the risk involved in specific experiments.

Indirect Harm: We also recognize that the field of network measurement — for the most part — focuses on understanding *systems* and not directly assessing *people*. Therefore, any impact to people is a side effect of our measurements. While we must grapple with the ethics of harm caused by our measurements regardless of whether the harm is direct or indirect, the nature of the harm can sometimes dictate the manner in which we cope.

Potential Harm: Next we note that most often our work does not cause harm, but rather only sets up the possibility of harm. That is, additional events or factors beyond our measurements must happen or exist for actual harm to be inflicted. Again, this does not absolve us from understanding the ethics involved, but does speak to how we may manage the risk involved in conducting a particular experiment.

We believe that while fuzzy, the above aspects of "harm" offer the broad contours of the issues with which researchers must grapple. Further, we do not believe there is some one-size-fits-all way to manage harm and we allow for honest disagreement among researchers about when potential and indirect harm rises to the level of making an experiment problematic. For instance, in the context of the example above about probes causing slight vs. serious harm, we discussed between ourselves whether periods of high-rate transmissions could be made short enough to reasonably be felt to avoid potential harm. We agreed it was possible, but disagreed about when the experiment transitioned from slight harm to serious harm.

4. COLLECTING DATA

Strictly speaking, active measurements have the potential to inflict direct and tangible harm. Passive measurements, by their nature, are simply recordings of observations and in no way directly change—benignly or harmfully—the operation of the network. Likewise, downloading and (re-)using a shared dataset does not alter the operation of the network—even if collecting it in the first place did. This latter brings up thorny issues of the use of so-called "found data". For instance, consider the the Carna botnet [4]. The botnet consisted of customer devices with guessable passwords that allowed illicit access, which in turn was used to take measurements which were publicly released. Clearly if a paper submission's methodology section read "we first compromised a set of customer devices" the paper would likely be rejected as unethical (and probably illegal!). However, if instead, one simply downloaded this data—causing no further harm to the customer devices or their users—is it ethical to use as part of one's research?

On the one hand, a researcher can make the case that any harm done by collecting the data has already transpired and therefore by simply downloading and using the data the researcher is in fact causing no harm. Further, if the data can provide insights into the network then perhaps we can view this as making the best of a bad situation. Alternately, we could view the use of such data as a moral hazard.

This issue is an open one in the medical community (cf. [14]). We will need to find our answers in the measurement community. There are likely different answers for different situations. For instance, a public dataset that was obtained by unethical means (e.g., the recent Ashley Madison dataset) may be viewed differently than a non-public dataset that happens to have been leaked to a researcher. We may view the first case as less problematic because of the reach of the data release, whereas in the latter case we may decide that the researcher is more culpable because, if not for their work, less would be known about the (potentially harmful) dataset. We encourage researchers to be thoughtful about the ethical issues related to the sources of their data.

5. STORING DATA

The measurement community generally encourages the preservation of measurement data to facilitate (i) revisiting the data in response to questions or concerns during the initial work, (ii) to look at new research questions later or (iii) historical comparisons. Furthermore, the community encourages researchers to make their data public, to better enable access to other researchers (cf. DatCat.org, an NSF sponsored repository of measurement data).

Preserving and publishing measurement data raises a number of ethical issues. We will highlight two.

First, how does a researcher determine if a dataset ethically can be made public? There are plenty of examples of successful deanonymization of data [15], and as the discussion in § 2.2.3 shows, our ability to extract information from seemingly innocuous data continues to improve. As an example, datasets published in the 1980s and early 1990s could likely be mined for passwords using packet timing algorithms published in 2001 [17].[3]

Second, if the data cannot be made public, but is retained, what safeguards are required to avoid accidental disclosure? For instance, should we expect all data stored on removable media to be encrypted? Should it be encrypted on non-removable disks? Do the rules vary according to the perceived sensitivity of the data?

We do not believe it is reasonable to expect researchers to anticipate all future analysis advances. However, we believe it reasonable to expect researchers to understand how current techniques could exploit their measurement data and to provide appropriate safeguards.

6. ON THE LIMITATIONS OF CONSENT

[2] Of course, we have experience with complaints about these sorts of probes, which indicates that some people do in fact view them as harmful.

[3] One risk is that users from the 1980s and 1990s who are still active today may still pick passwords in similar ways

One traditional way to deal with ethical issues that arise in an experiment is to require (informed) consent from the participants. This approach allows the people who could be potentially harmed by an experiment to weigh the possible harms against the possible benefits and to directly decide whether to participate. In some cases, Internet measurement can (and does) leverage consent. For instance, the Netalyzr measurement platform [10] aims to assess a user's Internet connection by providing a web page that the user must purposefully access. Further, the web page spells out what will happen and requires the user to explicitly start the measurements — hence consenting.

The Netalyzr situation is akin to the consent model in more traditional areas (e.g., medicine) and works well. However, in other settings obtaining informed consent for large-scale Internet measurements is significantly more difficult. Consider a study of end-user networks that uses a different methodology than that of Netalyzr. In [7], researchers use various tests to probe IP addresses they believe to represent home networks unbeknownst to the users. This provides a large-scale dataset that Netalyzr cannot match, but without the consent of the potentially affected people.[4]

Consent in Internet-scale measurements is difficult for two reasons. First, unlike, e.g., medical experiments, it is often unclear who is being measured and affected by Internet measurements. Further, even if the impacted actors could be identified, the logistics of obtaining consent range from significantly difficult to impossible.

While in more traditional areas of experimentation involving humans proxy consent is generally not allowed, in network measurements we lean on this mechanism. For instance, network measurements taken on a university campus typically seek consent from the university. However, probes sent off-campus may impact third parties with no connection to the university. Therefore, even while proxy consent can foster useful review to help identify and mitigate ethical issues, all possibly affected users are not covered directly or by some advocate.

In summary, there are cases where Internet measurements can leverage consent and we encourage researchers to do so in these situations. However, direct consent is not possible in most Internet measurements and therefore we as a community need to cope with ethical challenges without relying on consent.

7. PROPOSAL: AN ETHICS SECTION

As a community we lack norms or examples to guide researchers. Our position is twofold: (i) as a community we are not able to prescribe ethical norms for researchers to follow and (ii) therefore the best starting approach is to expose ethical thinking through a published "ethical considerations" section in all empirically-based papers. This approach serves three major goals:

- While some researchers are currently careful to understand the ethical issues surrounding their work, this care is not universal. Therefore, the first goal of an "ethical considerations" section is to force authors to publicly examine the ethical implications of their own work.

- Rather than counting on PCs and editors to impute the ethical foundations on which a piece of work rests, an "ethical considerations" section will give explicit voice to these issues. Reviewers will be able to directly evaluate the stated ethical implications of a piece of work and give concrete feedback to the authors, grounded in the authors' own words.

- Create public examples of good ethics. Currently ethics sections are not usually required by conferences, and if they are, are typically addenda to the paper seen by the PC and are not published.

We believe public ethics sections in papers will foster a conversation within the community, based on published exemplars, which will lead us towards norms.

We sketch four strawman questions authors should answer in an "ethical considerations" section. We aimed for a short list of questions—believing that capturing 80% of the ethics issues was better than a longer list that was still not exhaustive.

1. *For datasets directly collected by the author(s), could the collection of the data in this study be reasonably expected to cause tangible harm to any person's well-being? If so, discuss measures taken to mitigate the risk of harm.*

2. *For datasets not directly collected by the author(s), is there an ethical discussion of the collection elsewhere? If so, provide a citation. If not, the paper should include a discussion of the ethics involved in both collecting and using the data—beyond simply noting that no additional data collection harm occurs in re-using the data. This is especially important for non-public datasets.*

3. *Using current techniques, can the data used in this study reveal private or confidential information about individuals? If so, discuss measures taken to keep the data protected from inappropriate disclosure or misuse.*

4. *Please discuss additional ethical issues specific to the work that are not explicitly covered by one of the above questions.*

These questions intentionally do not address several items:

- There is no suggestion of when it might be appropriate to consult an Institutional Review Board (IRB) or similar body. These institutional bodies' involvement (or decision not to get involved) are not a substitute for the community's ethical review.

- We do not attempt to assess the ethics of the research *result*. Researchers are committed to advancing knowledge and in our view, that includes publishing results and techniques that may, if used unethically, cause tangible harm.

Furthermore, making ethics a core part of measurement papers will create some new challenges for reviewers and program committees, including:

- Review forms likely will have to be updated to ask reviewers to discuss the strengths and weaknesses of the ethics section.

- Mechanisms will be needed to help reviewers to evaluate ethics. Possible mechanisms include ethics guidelines from the program chair, ethics training, or simply an ethics teleconference at the start of the reviewing period. Over time, we hope prior published papers will help this process.

- Program committees will need to develop a clear philosophy on when papers are rejected based on ethical considerations and when papers with ethical gaps can be accepted, subject to revision. The issues surrounding collection (discussed in section § 4) will come up and program committees will need to find the community's answer(s).

[4] Note: Our point is about the difficulty of getting consent, *not an ethics criticism*. The authors of [7] properly sought to minimize the possible harm to the users and described their efforts in their paper.

Finally, what does it mean to reject a research paper on ethical grounds? While some papers may be resurrected by revising the work to not use an objectionable dataset, often the rejection will mean that the measurements used to support the paper's research results may have caused harm. That determination may raise a suite of questions about how to mitigate the harm and how to prevent such events in the future.

8. CONCLUSIONS

We present a strawman suggestion that authors of measurement papers include a (short) ethics section in their published paper. We believe doing so forces the surfacing of ethics issues surrounding individual measurement studies in a way that allows PCs to evaluate the ethics of a measurement experiment, and allows the broader community to move towards a communal ethical foundation.

Acknowledgments

This article benefits from many conversations with too many colleagues to name individually. Our thanks to all of them. Bendert Zevenbergen has organized multiple ethics discussions in the past year where this work has been usefully discussed. This work is funded in part by NSF grant CNS-1237265.

9. REFERENCES

[1] 18 US Code, Section 3121: General Prohibition on Pen Register and Trap and Trace Device Use; Exception.

[2] *Software Engineering Code of Ethics and Professional Practice; Version 5.2.* 1999.

[3] D. Bertsimas and L. D. Servi. Deducing Queueing From Transactional Data: The Queue Inference Engine Revisited. *Operations Research*, 40:217–228, 1992.

[4] C. Botnet. Internet Census 2012 , 2012. `http://internetcensus2012.bitbucket.org/paper.html`.

[5] S. Burnett and N. Feamster. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *Proceedings of ACM SIGCOMM 2015*, SIGCOMM '15, New York, NY, USA, 2015. ACM.

[6] D. Cousins, C. Partridge, K. Bongiovani, A. W. Jackson, R. Krishnan, T. Saxena, and W. T. Strayer. Understanding Encrypted Networks Through Signal and Systems Analysis of Traffic Timing. *Proc. 2003 IEEE Aerospace Conference*, Mar. 2003.

[7] M. Dischinger, A. Haeberlen, K. P. Gummadi, and S. Saroiu. Characterizing Residential Broadband Networks. In *ACM SIGCOMM/USENIX Internet Measurement Conference*, Oct. 2007.

[8] D. Dittrich and E. Kenneally. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research.* U.S. Department of Homeland Security, Aug. 2012.

[9] S. Floyd and V. Jacobson. The Synchronization of Periodic Routing Messages. In *Conference Proceedings on Communications Architectures, Protocols and Applications*, SIGCOMM '93, pages 33–44, New York, NY, USA, 1993. ACM.

[10] C. Kreibich, N. Weaver, B. Nechaev, and V. Paxson. Netalyzr: Illuminating The Edge Network. In *Proceedings of the ACM Internet Measurement Conference (IMC)*, pages 246–259, Melbourne, Australia, November 2010.

[11] R. C. Larson. The Queue Inference Engine: Deducing Queue Statistics From Transactional Data. *Management Science*, 36:586–601, May 1990.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Trans. Netw.*, 2(1):1–15, Feb. 1994.

[13] D. L. Mills. Internet Time Synchronization: The Network Time Protocol. *IEEE Transactions on Communications*, COM-39(10):1482–1493, Oct. 1991.

[14] P. Mostow. Like Building on Top of Auschwitz: On the Symbolic Meaning of Using Data From the Nazi Experiments, and On Non-use as a Form of Memorial. *Journal of Law and Religion*, 10:403–431, 1993.

[15] B. Schneier. Why 'Anonymous' Data Sometimes Isn't. *Wired*.

[16] D. C. Sicker, P. Ohm, and D. Grunwald. Legal Issues Surround Monitoring During Network Research. In *Proc. ACM Internet Measurement Conference*, Oct. 2007.

[17] D. X. Song, D. Wagner, and Z. Tian. Timing Analysis of Keystrokes and Timing Attacks on SSH. *USENIX Security*, 2001.

[18] R. Tomlinson. Selecting Sequence Numbers. In *Proc. ACM SIGCOMM/SIGOPS Interprocess Communications Workshop*, pages 11–23, Mar. 1975.

[19] S. (U.S.). Katz v. United States. 1967.

[20] W. Willinger and M. Roughan. Internet Topology Research Redux. In *ACM SIGCOMM eBook: Recent Advances in Networking*, 2013.

[21] C. V. Wright, L. Ballard, S. E. Coull, F. Monross, and G. M. Masson. Uncovering Spoken Phrases in Encrypted Voice over IP Conversations. *ACM Trans. Inf. Syst. Secur.*, 13(4), Dec. 2010.