Legal Issues Surrounding Monitoring During Network Research (Invited Paper)

Paul Ohm
School of Law
University of Colorado, Boulder
Paul.Ohm@colorado.edu

Douglas Sicker, Dirk Grunwald
Dept. of Computer Science
University of Colorado, Boulder
Douglas.Sicker@colorado.edu,
Dirk.Gruwald@cs.colorado.edu

ABSTRACT

This work was motivated by a discussion that two of the coauthors (computer science professors) had with the other coauthor (a law professor and a former computer crime Trial Attorney at the U.S. Department of Justice), in which it was pointed out that some of the network measurements that the computer scientists were thinking of making might potentially violate Federal laws.

Several Federal laws prohibit or restrict network monitoring and the sharing of records of network activity. These laws are designed to protect online privacy. They apply both to private parties and government agents, although the details vary depending on who is doing the monitoring. The most important thing to note is that none of these laws contain any specific exceptions or safe harbors for scientific or academic research. The laws are complex, but they follow a basic pattern. First, certain types of network monitoring and data access are prohibited. People who violate the prohibitions may be sued by the people whose privacy they invade and potentially prosecuted and convicted of federal crimes (i.e., misdemeanor and felony convictions).

In this paper, we will examine these laws and consider what they might mean for the network measurement community. Although we focus on U.S. Federal Law, we also highlight general trends and approaches in state and international laws that impact network researchers. We will examine the steps commonly taken in prior research in network measurement to respect user privacy, and we will compare those approaches to the evolving legal rules. We will also consider whether legislative reform is needed, describe steps that researchers might take when pursuing such work in light of the legal rules, and propose future technical and policy-related steps the community can take to focus more attention on user privacy.

IMC'07, October 24-26, 2007, San Diego, California, USA.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Monitoring; K.5 [**Legal Aspects of Computing**]: General

General Terms

Measurement, Legal Aspects

Keywords

Privacy, Legal, Monitoring, Law, Wiretap

1. INTRODUCTION

Research in the area of network measurement often requires the monitoring of actual network traffic. Many of the advances in this field are developed or confirmed only when such actual traffic is included. Privacy is often a paramount concern, and academic researchers tend to think a little or a lot about how to structure their research to minimize intrusions to privacy. Usually, these considerations are informed by a murky stew of social norms, rules of thumb, learned wisdom, and common sense about where to draw the blurry line between "private" and "public."

There is another ingredient in this stew, another source of rules that is often thought about but in the sketchiest terms: the Law. Most network researchers know that local, state, and national laws restrict some kinds of network monitoring and access to online records. But the law is poorly understood by many in the community, and there is much misinformation. Instead of clear and accurate rules, a vague understanding of the law is added to the stew. Dip a ladle in, and you can pull out the following rules of thumb, which are reassuringly clear and concise but not necessarily accurate. How many of these rules of the thumb have you heard? How many have you used to justify the legality of your research?

- "*It's my network, so I can do whatever I want.*"
- "*The network wiretapping laws have an exception for academic research.*"
- "*Packet sniffing is legal so long as you filter out data after the 48th (or 96th or 128th) byte.*"
- "*Capturing content may be illegal, but capturing non-content is fine.*"
- "*We're not breaking the law because we've anonymized the data.*"

- "*Data sent over a wireless network is available to the public, so capturing it is legal.*"

The goal of this paper is partly to educate. In section 2, we provide a primer for how law governs network monitoring and the disclosure of stored records, and we describe exceptions in the laws which sometimes immunize academic research. Our focus will be on the Federal law of the United States, although the discussion will also apply to the laws of states and other countries.

In section 3, we report the results of a literature review of past papers presented at the Internet Measurement Conference (IMC) 2005 [1] and 2006 [2]. From this survey, we identify a recurring set of strategies that researchers use to protect privacy. In conjunction with section 2, we conclude that steps taken to protect privacy in many of these papers may fall short of legal expectations. This part of the paper is not meant to point fingers. In fact, several of the papers that we have authored, fall short of clear legal compliance. We merely want to show the disconnect between the law and current academic practice.

Next, in section 4, we present some strategies for minimizing the risk of liability. Unfortunately, because the Law can be so vague, our proscriptions will probably leave many unsatisfied in two ways: first, compliance with the law might constrain measurement methodologies in ways completely at odds with the goals of most research; and second, while many of the rules will "minimize exposure," few will reduce the risk of liability to zero.

After noting some potential concerns, we ask whether there is something that can and should be done about the disconnect between law and research. First, in section 5, we offer some tentative proposals about how Congress might change these laws to allow for legitimate, academic research. The discussion is only tentative, because we conclude that there is no magic bullet fix that will sanction all legitimate research without creating a loophole for harmful, illegitimate behavior. Finally, in section 6, we talk about other steps the community of researchers can take to help make research and the law more consistent with one another.

Our intent with this paper is to educate and to begin a conversation within the community about user privacy and the law. We do not intend to claim that any member of the community has violated the law in the past, nor do we mean to point the finger at our community generally. In fact, it is important to underscore that we are not even in the position to make such judgments: it would be foolhardy to attempt to judge any past behavior without a detailed understanding of everything that occurred in the past research. Our only interest in the past is to help us begin a future-looking discussion about how we can conduct network monitoring that neither violates privacy nor hamstrings our research goals.

2. LEGAL BACKGROUND

Several Federal laws prohibit or restrict network monitoring and the sharing of records of network activity such as logfiles. These laws are designed to protect online privacy. They apply both to private parties and government agents, although the details vary depending on who is doing the monitoring. None of these laws contain any *specific* exceptions or safe harbors for scientific or academic research.

We focus primarily on the Federal law of the United States, and we wanted to explain why we chose this scope. First, these laws govern any research that occurs within the United States, so we expect our discussion to apply directly to a huge number of network researchers. Second, although most states of the U.S. have separate state laws that might govern network measurement research, many of those laws are modeled directly on the federal law [3].

Third, researchers in other countries can learn something from this paper because many international laws are similar to the U.S. Federal law. This is sometimes the product of outright copying and also because laws have been harmonized through treaty and convention [4].

Finally, even short of copying or treaty, legislators worldwide who have tried to regulate network monitoring have often taken similar approaches following some core general principles–consent, provider protection, anonymization–which are discussed in this paper. The odds are good that the legal principles described here apply to you no matter where you conduct your research, at some level of generality.

One last prefatory note: although Federal law is changed frequently, the law of network monitoring is quite stable. Some minor details change, but the specific features of the law discussed below have not significantly changed–either through legislative amendment or court interpretation–in over a decade. Accordingly, the following discussion will likely be accurate and relevant for a long time.

The Federal laws are complex, but they follow a basic pattern. First, certain types of network monitoring and data access are prohibited, although there are many complicated exceptions. Violations may lead to both civil and criminal liability. Simplifying things a bit, one set of laws applies to real-time monitoring–i.e. packet sniffing and the other to access to stored data. Let us consider each, in turn:

2.1 Real-Time Monitoring of Content: Wiretap

Packet sniffers are regulated by Federal Law. In particular, two laws originally designed to govern telephone monitoring apply, the Wiretap Act and the Pen Register and Trap and Trace Act. The Wiretap Act applies to monitoring of the content of communications (e.g., IM conversations; e-mail message bodies; and VoIP conversations) and the Pen Register and Trap and Trace Act applies to monitoring of *non-content headers* (e.g., most email, HTTP, and IP headers). Although there are some differences in the two laws, they are very similar. In this subsection, we will focus on Wiretap, content monitoring. In the next subsection, we will briefly highlight the Pen Register and Trap and Trace Act.

The Federal Wiretap Act, originally enacted in 1968 to protect against phone wiretapping and amended in 1986 to cover computer network communications, states a simple prohibition: thou shalt not intercept the contents of communications (See 18 U.S.C.§2511(1) [5]). Violations can result in civil and criminal penalties. The worst offenses may be investigated by the FBI, Secret Service, DEA, and IRS as felony prosecutions.

Congress has softened the reach of the law through a number of exceptions, some of which might apply in a research setting. There are two that are particularly important: (1) the provider protection exception and (2) consent.

The provider protection exception is the primary reason professional systems administrators can legally do their jobs

(see 18 U.S.C. §2511(2)(a)(i) [5]). The exception is easy to state but harder to define:

> *. . . network administrators can monitor a service as a necessary incident . . . to the protection of the rights or property of the provider of that service.*

In other words, monitoring packets to look for an intruder, sniff out a virus, or locate a bot, is probably justified as necessary to protect the network.

Courts have generally found that the *protection motive* need not be the only motive. So, if a sysadmin monitors both to protect the network and to help the police track down the culprit, the monitoring is still legal under this exception. (However, this exception is probably no longer available once the police begin to orchestrate the monitoring.) So a network security researcher may be able to justify research-related monitoring if there is some active protection goal in mind. The protection goal should be specific, however, and an argument that "*the results of my research will protect many of the world's networks*", probably will not fly, although no courts have expressed an opinion about that particular argument.

Even monitoring to protect a network has its limits. Courts will examine the scope of the monitoring to see if it is related to the goal. In legalese, Courts will verify that there is a "substantial nexus" between the monitoring and the threat. Capturing every single packet on a network may seem like an effective way to hunt down a virus, but a court could conclude that this is too much monitoring for the purpose. Of course, to help them make this judgment courts are likely to look at things like industry practice.

The great virtue of the provider protection exception is that it is very easy to assert; so long as you are acting on behalf of a provider and protecting the provider's rights or property, your monitoring may be covered. The problem with the exception is vagueness. What is a "right"? What is "property"? How should "substantial nexus" be defined? Courts have yet to provide clear answers to these questions.

In contrast, the other important exception, the *consent exception*, can provide certain, absolute protection, but it is often very difficult or even impossible to use in the research setting. Simply stated, network monitoring is not illegal if at least one party to the communication has consented to be monitored (See 18 U.S.C. §2511(2)(d) [5]). This is not the same thing as saying that only one person in your entire packet trace need have consented, because a typical trace easily contains tens of thousands (or more) separate "communications." If a sniffer is capturing an instant messaging conversation, that conversation is a communication and to fall within the exception, at least one of the participants in the IM chat needs to have consented. If SMTP traffic is captured, then the "parties" are the sender, the recipient, and possibly the various SMTP servers involved. If IM and SMTP packets flow alongside one another into the sniffer, they will be considered distinct communications for purposes of consent. Thus, for example, it is inconceivable that the consent exception can be used to justify monitoring backbone traffic.

Note that sometimes the researcher himself may be a consenting party. For example, recording IRC conversations in which the researcher participates is almost certainly legal. Also, keeping a typical web server's access log is legal because the web server admin (or, more metaphysically, the webserver itself) is a consenting party to the communication. These acts, however, might still violate the law in the twelve states that require the consent of all parties to the communication. As of 2003, the twelve states that required the consent of all parties to a communication were California, Connecticut, Florida, Illinois, Maryland, Massachusetts, Michigan, Montana, Nevada, New Hampshire, Pennsylvania and Washington [6]. For example, a judge in New Hampshire ruled that a police officer violated the state's all-party-consent wiretap law by using session capture software to make a video of an AOL chat room conversation in which he posed as a 14-year-old girl to lure child predators [7].

Assuming you are not a party, how do you get the consent of your monitoring subjects? The clearest form of explicit consent is a signed sheet of paper, such as a network monitoring policy or terms of service contract. For example, on a college campus where every user of the network has signed a document that says simply, "*I consent to have my network traffic monitored for research purposes*," monitoring these users may be legal. But even explicit consent can be ambiguous. If the signed statement says, "*In order to help the University combat fraud, I consent to be monitored*," then the consent probably does not apply to monitoring unrelated to fraud detection and prevention. Similarly, a statement that says, "*IT staff may monitor my network activity*," will probably not apply to non-IT employees, such as researchers in the computer science department. Although signed pieces of paper are best, many courts have found valid consent when a user has clicked an "I agree" button.

Finally, consent may be implied. Implied consent means, "*based on his behavior, it seems like he agreed to be monitored.*" For example, if an employer repeatedly tells an employee that his network activity may be monitored, and if he doesn't protest, he has probably consented [8]. Similarly, if a system displays a banner that reads, "*by using this system, you agree to be monitored*," then use likely equals consent [9].

But implied consent does not mean, "any reasonable person should have known that they would be monitored." Even if the average Internet user assumes that their communications are being monitored, their continued use in the face of such worry does not equal consent.

2.2 Real-Time Monitoring of Non-Content: Pen Register and Trap and Trace

The Wiretap Act applies only to the monitoring of the *content* of communications. Prior to 2001, the monitoring of *non-content*, header data appeared unregulated under Federal law. The old version of the law applied specifically to "numbers dialed" and "originating number[s]" relating to telephone calls.

In 2001, in response to the 9/11 attacks, Congress passed the infamous USA PATRIOT Act. Along with many other changes, this law amended the Pen Register[1] and Trap and Trace Act to apply it to "device[s] or process[es]" which "record or decode" or "capture" the "dialing, routing, ad-

[1]The term "Pen Register" gives you a sense of the historical nature of the law. A "pen register" is the component of a telegraph system that records a received message. That meaning evolved to denote a device that records all calls originated from a particular phone. A "trace and trap" device records the number of all *in-coming* calls on a particular phone line.

dressing, or signaling information associated with electronic communications. (See 18 U.S.C. §3127 [10]).

The fact that for fifteen years this law appeared not to apply to non-content monitoring may be the source of many of the misconceptions about the law discussed at the beginning of this paper. Before the USA PATRIOT Act, there was a good argument that non-content monitoring was legal.[2] Whether or not those arguments were valid at some point in the past, under current law it is clear that Congress has regulated non-content monitoring on computer networks.

As always, there are ambiguities in the definitions: for example, is a URL "dialing, routing, addressing, and signaling" information falling within the Pen Register and Trap and Trace Act, or is it "content" falling within the Wiretap Act? But there is no ambiguity that IP address information is "dialing, routing, addressing, [or] signaling" information and thus covered.

There are many differences between the Pen Register and Trap and Trace Act and the Wiretap Act, but none of these differences will matter much to the network monitoring researcher. For example, capturing non-content, header information is a misdeamor, while intercepting content is a felony; the important thing is that both acts are prohibited. Like the Wiretap Act, the Pen Register and Trap and Trace Act allows exceptions for provider monitoring and consent, but the limits discussed above on using those exceptions to justify network monitoring research apply in this context, as well.

2.3 Sharing Stored Records of Network Activity

Another common research activity that is governed by federal law is the sharing of particular types of stored records of online activity. These activities are governed by the Electronic Communications Privacy Act [11, 12, 13]. Again, important distinctions are made depending on whether the records include content (e.g. full e-mail messages) or non-content (e.g. webserver access logs).

The basic rule is, again, a prohibition: some network service providers are prohibited from giving content and non-content stored records to others. There are some very important, broad exceptions to this rule. First, *non-public* providers don't fall under the prohibition. So the webmaster for a private company can legally hand over access logs to network researchers. Second, any provider can hand over *non-content* records to anyone except the government. So, traffic data, saved headers (but only the ones that do not contain any content), and other logfiles can be shared freely, but not with the government. None of these exceptions, however, permits public providers (like AOL) to share stored content like e-mail messages.

This raises another curiosity in the law: because we work for the University of Colorado, are we and other re- searchers at public institutions considered "government entities?" If we are, then it would be illegal for some sysadmins to share logfiles with us. This result seems highly irrational, but it is not an improbable reading of the law. There is even a remote possibility that a researcher at a government-run lab is a government actor for purposes of the Fourth Amendment, which primarily governs police searches and seizures. The Fourth Amendment has been held to apply, for example, to government employers surveilling their employees and public school teachers searching student lockers.

As with the Wiretap Act, there are other exceptions to these prohibitions. Both the consent and protection of rights and property exceptions described above permit the disclosure of stored communications, as well.

[2]For reasons outside the scope of this paper, during many of those years, despite the seemingly technology-specific language in this law, the Department of Justice interpreted it to apply to non-content, header monitoring on computer networks, as well as to phones.

3. LITERATURE SURVEY: PRIVACY PROTECTION IN PAST RESEARCH

In order to assess the practices for current networking research, we surveyed a total of 57 papers published in the IMC 2005 and 2006. We prepared a "coding sheet" to record specific facets of how the individual research papers collected and used any data or traces collected for the study. The coding sheet contained 15 features of the traces used in the papers. These included:

- The approximate "scale" of the network (such as Tier-1 network, etc),
- Whether the network was accessible to the general public or was a private corporate or internal network that requires special access,
- The approximate number of people affected by any tracing activity,
- Whether the data was recorded and aggregated in real time or stored and analyzed later,
- If the data was stored and analyzed,
  - How much of the packet was recorded (headers or full packets)
  - How were "headers" indicated or demarked
  - Was the data anonymized and if so, how
  - Was the data "filtered" and then saved to a recording media or saved and then filtered,
  - What, if anything, was done with the data remaining after filtering,
- Was the data disclosed to other parties,
- Was there a possible "protection motive", where data was gathered to protect a network resource or assets,
- What tools were used in the data collection,
- and, lastly, was the data collected from a wireless network.

Following this literature survey, we found that several common patterns emerged and that rather than present an exhaustive discussion of the coding, it would be more instructive to look at the common cases and discuss exceptions. Table 1 provides a summary of the coding sheet. Some papers contribute to multiple counts because they use multiple data sources. The original coding distinguished between information explicitly provided by the articles, including information that was not appropriate or not indicated in the paper.

Public/Private Network	Network Level	Information Captured	Number Papers
Public	Tier-1	Headers	9
Public	External corporate or University system or web server	Headers	12
Public	Other (e.g. IETF meeting, p2p network)	Headers	5
Public	Tier-1	Full Packets	1
Public	External	Full Packets	6
Public	Other	Full Packets	
Private	Tier-1	Headers	–
Private	Internal University, Corporate / Enterprise system	Headers	1
Private	Other (cable modem net, etc)	Headers	–
Private	Tier-1	Full	–
Private	Internal	Full	2
Private	Other	Full	1

Table 1: Summary of literature coding for papers published in IMC 2005 and IMC 2006.

In this reduced survey, we have assumed that papers that did not indicate a source should be resolved in a way that assumes the researchers recorded and accessed the data in the most legally compatible fashion (e.g. did not store the data, did not log full packets, etc). We only coded and included papers that use new sources or traces; some studies used standardized traces from network monitoring organizations.

3.1 General Results

We found that most traces involve measurement of "public" environments, or networks or services that were accessible to the general public or Internet consumers. Examples of these networks included customer-facing corporate web servers, public spaces in Universities (e.g. aggregation points for campus networks rather than a specific subunit or department) or ISP links servicing multiple aggregates. This result is not surprising because "Internet" researchers are obviously interested in the Internet rather than Intranet properties. However, monitoring of these public networks raises many of the issues discussed in Section 2. Although a corporate web server may provide a privacy statement, few consumers read those privacy policies and the policies would need to describe that the recorded information (including e.g. the consumers IP address and packet contents) may be distributed. This consent is typically easier to establish in a private organization or enterprise where such monitoring may be a condition of employment.

Almost all of the papers collected data (either headers or full packets) and then filtered that data for specific content. For example, researchers might use a full packet data capture from a cable modem network provided by an ISP and then filter out just the headers for bittorrent or p2p traffic. As discussed in Section 2, the legal distinction of recording full packets is significant for the law.

Only one of the papers using new sources or traces explicitly discussed anonymizing the data. That anonymization was performed by "randomizing" the data using a salt and a cryptographic hash; given the hash and the knowledge of the "salt" value, this trace could be inverted using brute force methods. Other papers used standardized traces (discussed later) that are typically anonymized.

Most of the studies using traces from public networks involved multiple institutions; for example, the data might be collected by a Tier-1 ISP but paper co-authors are from Universities or research labs. Our assumption was that in such papers, some aspect of the data was disclosed between organizations with the implications of Section 2.2 discussed above. None of the papers discussed how the data was shared or what restrictions were placed on the data.

Very few of the papers discussed the decision process used to distinguish "non-content" from "content". Typically, fixed sized header lengths or NetFlow mechanisms were used to collect such data. Again, this issue can be important from a legal perspective. It is worth pointing out that even if an author had a clear understanding of what constitutes proper handling of these privacy data, it is not likely that they would want to "waste" space in the paper, given that it is already difficult to fit the main concepts while adhering to the page limits.

3.2 Commonly Reused Traces

A number of shared traces are commonly used among the network community. Among the many well known shared datasets are the PMA NLANR traces, which provide passive header trace data, and the CAIDA skitter measurements. These traces offer researchers with a ready source of data to serve as inputs to emulate and model network conditions and as raw data for statistical analysis of network behavior. Of course it is important that the reuse of these traces do not introduce biasing, due to the characteristics of the given sample data. In each of these datasets care has been taken to provide anonymization. For example, CAIDA applies Crypto-Pan [14, 15] to anonymize data taken from a tier 1 ISP. This is a prefix preserving technique, which maintains a meaningful mapping of the anonymized data. Pang *et.al.* [16] examined the difficulties of anonymizing traces and described the steps taken to develop a anonymization policy and validate its correctness. Part of this work included the

development of a tool, `tcpkpub`, for the task of anonymizing trace data. A variety of other tools also exist for such anonymization.

4. STRATEGIES FOR PROTECTING USER PRIVACY

We now present some strategies for protecting user privacy, particularly in light of our discussion of the law. Unfortunately, with all of the ambiguities and unanswered questions in the law, it is impossible to provide a concise and precise set of rules. Furthermore, some of the recommendations which follow are likely not useful to researchers because they can't be complied with or would defeat the goals of the research. As an impractical example, one surefire way to avoid liability entirely is not to collect any data, which is very clear legal advice, but a research non-starter. As you read this, remember that we are not your lawyers (indeed, two of us aren't lawyers at all!), so please consult with University or Corporate counsel before proceeding.

4.1 Get Consent

Consent is the clearest exception in the law. Obtaining clear, explicit consent from every user on the network is the best way to minimize legal exposure. Consent is easier to obtain, of course, with networks of small user populations. Therefore, limiting monitoring to a single research-group (best) or department (still good) may limit liability, so long as every single monitored user has signed or clicked on a consent form. Remember that the mere fact that a user has signed is not good enough; the language of the form must cover the type of monitoring undertaken. Also, if monitoring pursuant to consent, be sure to have controls and checks in place to ensure that you are monitoring only the users who have consented and that data from other users isn't leaking across a router or switch.

Note that researchers who work for commercial ISPs such as Cable companies and phone companies and web companies like Google and Yahoo! may already have the consent of their users. These kinds of providers tend to get expansive consent from all of their users to monitor without limits. In fact, complying with the law may be much easier for a Google researcher looking at the private data of his users than it is for Academic researchers on University networks or private researchers in closed Corporate networks engaged in the same activity.

4.2 Use Simulated Data

Another way to minimize liability is to use simulated data. There are a number of notable tools for producing simulated data, such as Harpoon [17] and Swing [18]. While these are useful tools, there are limits to the usefulness of simulated data, as will be discussed further in section 6.

4.3 Anonymization / Data Reduction / Minimization

As the literature survey suggests, when researchers talk at all about the steps they've taken to protect user privacy, they often mutter the magical incantation, "*we anonymized the data.*" There is a persistent belief that data collection is legal so long as the data is reduced, minimized, or otherwise pseudonymized or anonymized. This is sometimes true, and sometimes not.

Simply put, data reduction is not the be-all, end-all of legal compliance that many researchers believe it to be. There are at least three reasons. First, there are many possible approaches to anonymization, some better than others. One common technique is to map IP addresses to unique, scrubbed identifiers, permitting analysis across sessions but protecting against IP address attribution techniques. The pitfalls of this approach were demonstrated in August, 2006, when America Online released the search records of 650,000 users. Although IP addresses were removed from this data, by comparing the search query strings of certain "anonymized" IP address identifiers, journalists from the New York Times were able quickly to reveal that user 4417749 was Thelma Arnold of Lilburn, Georgia [19, 20].

Second, the words "anonymous" and "pseudonymous" do not appear (in any grammatical form) in the federal laws described in this paper. There are certainly reasonable arguments that anonymizing data may sometimes immunize behavior: for example, the laws described talk about restricting access to data associated with "contents," "communications," "users," "customers," "subscribers," or "sources." If prosecuted or sued, a researcher might successfully argue that he had thrown away so much data, that what was left no longer met the meaning of any of these critical words. But that's just an untested legal argument that turns in unpredictable ways on exactly how much was thrown away. A related misperception is that collecting content is illegal but capturing non-content traffic data is legal. While non-content collection doesn't violate the Wiretap Act, it might violate the Pen Register/Trap and Trace Act.

Third, even legally sufficient anonymization is no defense if clumsily executed. For example, keeping copies of the original, non-anoymized data is as bad, as far as the law is concerned, as not anonymizing at all. Similarly, doing the filtering/data reduction days or weeks after collection probably won't immunize the behavior completely, although it may reduce the cash verdict given to victims. In fact, some tools, such as the popular wireless monitoring tool kismet, can be easily misconfigured to filter the data displayed to the screen while saving full packet data to disk. If FBI agents find on your hard drive gigabytes of stored packets that you thought you had filtered away, they may have trouble believing you when you say that you didn't know they were there.

4.4 The Network Researcher's Motto: First, Do No Harm?

Even using this article as a guide, it is highly likely given the vagueness in the law that academic researchers will be forced to conduct research without knowing whether or not they are complying with the law. In those situations, it is important to structure monitoring with one pragmatic feature of the justice system in mind: laws like these are usually not invoked until and unless victims complain. If researchers collect and maintain data under controls that scrupulously protect the privacy of the people whose communications are intercepted, even if a law is technically violated, there may be no claims of harm and thus no repercussions. We are not saying that it is appropriate to break the law if nobody complains. But given the vagueness of these laws, researchers often won't be able to tell if they've crossed the line between

the permissible and the illegal, and in those cases, it is wise to do everything possible to avoid harming anybody.

In some sense, then, we're back where we started. We began by criticizing the informal rules of thumb that people mistakenly think will immunize network monitoring. We hope we have put to rest the misconception that rules like these offer clear, absolute immunity. Nevertheless, rules of thumb like the following are important if they minimize the risk of harm, and thus the risk of civil suit or criminal prosecution:

- Capture only the data you need.
- If IP addresses can be scrubbed, scrub them.
- If content or IP addresses must be stored, encrypt the data when you're not analyzing it.
- Restrict monitoring to the smallest network that satisfies research requirements. (e.g., don't use a backbone provider's OC-192 when a University LAN will do.)
- If filtering is used, be aware that some tools may store the unfiltered full packets to disk.

5. IS LEGISLATIVE REFORM NEEDED?

Congress, in the laudable pursuit of trying to protect online privacy, did not take into account legitimate research that can only be done through network monitoring or disclosure of stored data. Should Congress change the laws to accommodate legitimate Academic research? There are reasons to hope that they don't try.

First, it is very hard to draw the line between impermissible and permissible monitoring. At least two approaches are possible, neither satisfactory. The first approach is to exempt activities based on status. For example, academic researchers," or "professors," may be exempt. Obviously, this is both over-inclusive and under-inclusive.

Another approach is to try to build purpose or motive into the law: for example, "all monitoring done pursuant to federally-funded research is exempt from this law." This too is under-inclusive (what about non-federally-funded research?) and over-inclusive (just because it's federally-funded doesn't mean it should be legal). The other problem with tying illegality to purpose or motive is that the rules will be necessarily vague and hard to predict, and the resulting cases will be much more complex and difficult to investigate. If the line between legal and illegal monitoring is what was in your e-mail inbox or innermost thoughts, the FBI might feel obligated to prowl through both to decide if they have a case.

Second, Congress has a checkered history of trying to accommodate academic research. Most notably, consider the Digital Millennium Copyright Act's prohibitions on circumvention of DRM technology. Many academic researchers have complained that the law chills encryption research. Congress attempted to address this, but the result is a lesson in the difficulty of crafting these kinds of exceptions. Subsection 1201(g), the "Encryption Research" exception, includes 5 subparts and 14 sub-sub-parts. The law requires researchers to jump through many, many hoops before earning the exemption. Only some types of research are exempted, and researchers must attempt to get manufacturer authorization before proceeding. Even after passing through these hoops, the research must survive a multi-factor test that turns on, among other things, the use of the information derived and the status of the researcher (including gems like "legitimate course of study" and "appropriately trained or experienced").

Another example perhaps more closely analogous to the network monitoring situation are the Health Insurance Portability and Accountability Act, or HIPAA, privacy rules and their impact on medical research involving clinical studies. Under the law and Department of Health and Human Services regulations, such research can occur, but only subject to informed patient consent, and even then with many restrictions on the use and disclosure of the private information.

Two other examples are patent law, which has a weakly interpreted "research exception," and the "fundamental research" exception of the International Traffic in Arms Regulations and Export Administration Regulations (ITAR and EAR) that together limit export of certain materials, devices, and technical information (including software) to certain countries.

Although a detailed examination of these laws is beyond the scope of this work, suffice it to say that these exceptions are generally considered to be narrow, onerous, and complex. It is not easy to define an academic research exception for behaviors that policymakers generally find harmful.

A bit more optimistically, a change may not be necessary because of police discretion. Simply put, the police are extremely unlikely to investigate the vast majority of academic network monitoring that occurs. In fact, the only time the police would possibly get involved is when a victim steps forward and calls foul, claiming that the researcher crossed some line between desirable and undesirable behavior. It may be that we want the police to have the freedom to investigate the worst examples of those "line-crossing" cases, and an immunity for research might foil such investigations.

But the problem with relying on police discretion is that some people will not exercise their discretion wisely. Over-aggressive prosecutions and investigations happen, and broad laws can be used to support them.

6. WHAT CAN THE RESEARCH COMMUNITY DO?

Assuming a legislative fix won't happen soon, what else can the network measurement research community do? We have several proposals. First, the paper has referred repeatedly to the vague body of social norms that separate the permissible from the impermissible with network monitoring. We each seem to have our own intuitions for the general contours of these norms, but maybe we should try to come to a community-wide consensus. At the very least, we should proceed informally, by beginning to have conversations about what constitutes acceptable network monitoring. Although we probably won't agree on every detail, a starting point is to assess the traditional privacy-protecting strategies that are repeatedly used, as reflected in our literature survey. Once we begin to agree on those norms, we may need a formal effort to write down the rules, perhaps in an informational RFC. A codified understanding that reflects even rough consensus would be a useful tool to bring to Congress or to show to courts. It is important that these norms and rules are agreed upon from within our commu-

nity, rather than dictated to us by some outside court or agency.

Second, although consent to monitor is a clear, absolute method for avoiding liability, consent is very hard to obtain, at least for non-trivial population sizes. Perhaps we can develop protocols or applications for querying users for consent to monitor, along with response mechanisms which signal various levels of permissible monitoring.

Third, we should try out different approaches to privacy enforcement on smaller networks before the entire Internet. For example, research networks like Abilene, Internet2, or GENI, which are testbeds for new technology should also become testbeds for approaches to privacy. Maybe subparts of GENI can be declared completely "consentful," meaning every user on that subpart knows that their communications may be monitored. Or maybe Congress should be urged to pass explicit exceptions from state and federal monitoring law but only for a particular, small research network.

Fourth, we should support and pursue research into simulating network traffic. Obviously, network traffic simulation involves tradeoffs and raises doubts about the faithfulness of the simulation. Privacy concerns might justify greater attention on this kind of research.

Finally, our literature survey demonstrated the common practice of reusing well-known, historical packet traces. Under the wiretap act, if data is illegally intercepted, then every subsequent disclosure and use of that data is a separate violation (See 18 U.S.C. 2511(1)(c), (1)(d) [5]). We should insist that commonly reused packet traces be clearly marked to indicate the privacy protection measures used in the creation of the data set. To forward this goal, we can set up a standard language for describing various privacy-enhancing measures, along the lines of our coding categories from Section 3.

7. ACKNOWLEDGEMENTS

This project is partially funded by NSF Project #0435297, #0454404 and #0435452.

8. REFERENCES

[1] ACM SIGCOMM and USENIX, *Sixth Internet Measurement Conference*, 2005. Available online at `http://www.imconf.net/imc-2005/papers/program.html`.

[2] ACM SIGCOMM and USENIX, *Sixth Internet Measurement Conference*, 2006. Available online at `http://www.imconf.net/imc-2006/program.html`.

[3] C. H. Kennedy and P. Swire, "State wireless and electronic surveillance after september 11," *Hastings Law Journal*, vol. 54, no. 847, 2003. Appendix A.

[4] C. of Europe, "Convention on cybercrime budapest 23.xi.2001." Available as `http://conventions.coe.int/Treaty/EN/Treaties/HTML/185.htm`.

[5] "18 united states code § 2511." Available at `http://www4.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002511----000-.html`.

[6] *The First Amendment Handbook*. The Reporters Committee for Freedom of the Press, 2003. Available as `http://www.rcfp.org/handbook/c03p01.html`.

[7] M. Rasch, "Chat, copy, paste, prison," *SecurityFocus*, April 2004.

[8] Griggs-Ryan v. Smith, 904 F.2d 112 (1st Cir. 1990).

[9] U.S. v. Angevine, 281 F.3d 1130 (10th Cir. 2002).

[10] "18 united states code § 3127." Available at `http://www4.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00003127----000-.html`.

[11] "18 united states code § 2701." Available at `http://www4.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002701----000-.html`.

[12] "18 united states code § 2702." Available at `http://www4.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002702----000-.html`.

[13] "18 united states code § 2703." Available at `http://www4.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002703----000-.html`.

[14] J. Xu, J. Fan, M. Ammar, and S. Moon, "Prefixpreserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme," 2002.

[15] "Crypto-pan software," 2004. Available from `http://www.cc.gatech.edu/computing/Networking/projects/cryptopan/`.

[16] R. Pang, M. Allman, V. Paxson, and J. Lee, "The devil and packet trace anonymization," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 1, pp. 29–38, 2006.

[17] J. Sommers. and P. Barford, "Self-configuring network traffic generation," in *Fourth Internet Measurement Conference*, 2004.

[18] K. V. Vishwanath and A. Vahdat, "Realistic and responsive network traffic generation," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 4, pp. 111–122, 2006.

[19] M. Barbaro and T. Z. Jr., "A face is exposed for aol searcher number 4417749," *New York Times*, Aug 2006.

[20] C. Soghoian, "The problem of anonymous vanity searches," Jan 2007. Available at SSRN `http://ssrn.com/abstract=953673`.