High Fidelity, High Risk, High Reward: Using High-Fidelity Networking Data in Ethically Sound Research

Mohammad Taha Khan
taha@cs.uic.edu

Chris Kanich
ckanich@uic.edu

Department of Computer Science
University of Illinois at Chicago

ABSTRACT

Network tap data can provide researchers with access to every packet flowing into or out of an organization. However, building a sound ethical framework around using this data is a necessary task before the community can embrace this data source. Here we describe the ethical issues, present example use cases, and suggest strategies for creating a strong ethical footing for this research while maintaining some level of utility to the researchers.

1. INTRODUCTION

Network taps that have the capability to observe every bit going in or out of a large organization can be incredibly helpful for networking research. The human-level decisions and interactions of network users are becoming the focus of networking research, and thus anonymized header-only traces are no longer sufficient to answer these questions. At some institutions, the opportunity exists to access full fidelity network tap feeds in the name of improving the operation and security on the network. There are two potentially serious downsides to using this capability. Firstly, this invasion of privacy can in and of itself be harmful to users of the network, through bad actions by the researchers or a breach of stored data. How much harm this infringement could possibly cause, and whether its likelihood and damage are commensurate with the potential benefits of the research, is a matter which must be understood by the researchers, their community, and the human subjects research review board (IRB in the United States) so that they can properly evaluate the research. The second and possibly more serious concern is that the mere existence of this tap and its use for research (even when explicitly allowed by an AUP) may be damaging to network users in the organization for instance through "chilling effects" on the use of the network to perform sensitive communication.

Even though such chilling effects are possible, these taps likely already exist at some if not all major research organizations and their traffic is visible to employees in roles related to network operations. Allowing researchers to have similar access to this traffic may well present a great boon to networking research that focuses on human effects and interactions rather than network or computer interactions, but must be developed and presented in a way that takes full consideration of all stakeholders involved.

In this paper, we present a list of the main data sources available at such taps and the types of measurement and analysis which are possible. Using the framework developed in the Menlo Report [6], we address the concerns that stakeholders may have and summarize ethical guidelines and reporting considerations [1] which would be advised when dealing with tap data. By using specific examples, we elaborate on the in-context use of these guidelines. Finally, we suggest specific considerations for data access, in hopes of beginning a discussion which will enable the community to delineate research norms that are fair to research subjects without unduly stifling research progress in this incredibly important field.

2. RELATED WORK

With the advent of large-scale data driven research projects, many efforts in the networking research community aim to ensure that measurement data is accessed under the terms of an acceptable usage policy. In the past few years, researchers have focused on developing systems [2, 8], and methods [7] to allow entities to share and perform research on measurement data in a sound manner.

More recently, due to the essential need for measurement research to be conductible by investigators beyond those with direct access to data, NSF and Homeland Security have developed DatCat & Predict [5, 4], which are repositories to provide researchers regulated access to network data. As the aim of such studies is to minimize risk and maximize the research value of a certain dataset, researchers have developed methods for maintaining ethical access methods while maximizing the utility of the datasets. Sometimes research organizations and enterprises are reluctant to share data. Researchers have addressed this problem with novel techniques [3] that modify packet traces to ensure the anonymity of

shared network data. However, these methods provide anonymization on the network layer and only serve useful for purely technical analyses. In situations where the behavior of the human subjects is under investigation, it is nontrivial to anonymize data while maintaining any semblance of usefulness.

To establish a privacy-preserving environment in the research community, measurement conferences like *IMC*, *PAM* and *FOCI* also require authors to adhere to limits of use prescribed by the IRB.

The Menlo Report [6] and Allman and Paxson [1] provide specific recommendations on the usage and considerations of handling network related data. In this work we apply these guidelines to the use of high fidelity data obtained from network taps.

3. DATA SOURCES

This section describes the types of data sources commonly used for measurement research and how they relate to potentially sensitive information. The two main types of data we consider are data passively collected from a local vantage point, and data actively collected or requested from outside of the organization.

3.1 Enterprise and Institution Data

HTTP and DNS Logs: To date, a major portion of Internet traffic is transmitted over unencrypted HTTP streams. This allows datasets obtained from taps to provide full visibility into various characteristics and trends. When directly collected with an infrastructure like Bro, these logs can contain several useful fields such as timestamps, IPs, URIs, or even relevant cookies set or sent. For instance, The presence of timestamps enables studies of a temporal nature whereas IP and URI information are essential for looking into browsing trends. Within the tech industry, HTTP access logs have been used extensively for optimizing page design and content delivery, and passive collection of the same records could be used for similar tasks.

Middleware Data and Reports: This category includes data logs generally obtained from proxies, firewalls, and filtering products. In some cases, the aggregation of end user software reports such as malware logs also serve as a rich source of information. These types of datasets are usually affiliated with a business environment where a central system uniformly controls end hosts. This homogeneity, along with the fact that many if not all devices and software are owned and controlled by the enterprise, place research using this data on slightly stronger ethical footing. However, one must still consider employees as stakeholders in this case, and research should be done in such a way to minimize harm wherever possible.

3.2 Other Sources

Crawler Data: Crawling is an active mechanism of obtaining data from the web. Although, it is a systematic collection of publicly available data, the aggregation of information into a searchable format qualitatively changes the privacy invading properties of such a dataset. Crawling websites that relate to human subjects are commonly performed on social networks and in some cases, forums and pornographic websites.

Botnet/Honeypot Data: Botnets, comprised of end user computers, wireless devices, or network infrastructure like home routers, can also be leveraged as an active method for data collection. In some instances, researchers have set up online honeypots in the form of "fake" profiles and pages on social networks to collect user information. This activity is directly linked to exposure of subjects' private information when interacting with these research platforms.

Data From ISPs: ISPs provide the backbone infrastructure of the Internet. Similar to network taps, information obtained from these sources can provide deep insight into human level network usage trends. As opposed to enterprise/institution data, subjects of such a dataset are of a much more diverse nature, which enables analysis on a much broader scale. Such data collections are commonly used to identify viral trends pertaining to events that occur on a temporal scale. Examples include political and social events. Apart from trending online, these events have a prominent "offline" presence which greatly raises the concern of how human subjects are depicted as a part of such research.

Public Sniffing Probes: In the past, researchers have used mobile sniffing probes in public environments to collect data from Wi-Fi networks. Data characteristics of such a collection are similar to HTTP based logs, however, the nature of subjects under consideration can be of varying nature, depending on the vantage points of collection. While there are methods for sanitizing this data, again it maintains the utility at the network level while disallowing analysis "higher in the stack," for instance following users over extended periods of time or at different layers simultaneously, both where they are and what they are looking at on a mobile device. Although this collection technique is generic, the risk posed to subjects is of equal magnitude when compared to subjects whose data is specially obtained from an ISP, enterprise, or an institution.

4. ANALYSIS OPPORTUNITIES

In this section, we elaborate on how datasets like those described in section 3 can be used to perform various measurements and analyses.

Large Scale Measurement Analysis: This category encompasses a broad range of measurement techniques. We briefly explain each technique below.

- **Statistical/Trend Analysis:** This is a general

measurement technique that focuses on providing numbers and probabilities of a specific activity or trend within a dataset. Sometimes, researchers focus on answering a fundamental yes or no question from a relevant dataset.

- **Model-Based Analysis:** Other approaches use machine learning techniques to develop general models of various phenomena. Using relevant features from the data, the model can be used to predict future trends or scenarios that might arise later on. Mostly, trends are associated with user activity.
- **Performance Analysis:** Performance-based measurements are primarily concerned with analysis of the quality of a system or infrastructure associated with the dataset. A performance metric is defined by analysis of user activity but the research focus is on the systems as opposed to the human subjects.

User Profiling: In this type of study human subjects are the primary concern. Researchers focus on answering questions about various aspects of user activity present within a dataset. Some of the common analysis techniques include:

- **Online Activity:** Most network logs provide user level granularity. This allows the study of online activity which includes examples like browsing trends or e-commerce.
- **Behavioral Analysis:** This analysis is primarily concerned with looking into the online behavior of users. Although this is similar to online activity, the focus of such research is on a specific subset of users on an online social network or forum. Information for these kinds of studies is extracted using web crawlers or social network APIs.
- **Direct Interaction:** User profiling can also be performed through direct interaction with human subjects. Subject identifiers are extracted from within the dataset and evidence for research is built upon direct interactions. Examples include phone calls, emails, and social network connections. Interestingly, in this form of analysis, researchers are themselves exposed to a certain form of risk depending on the nature of subjects under consideration i.e. interaction with online drug dealers.

Cybercrime Analysis: The analysis of malicious activity and the quantification of cyber crime within an institution or enterprise is an important aspect of measurement research. This is usually performed on HTTP logs, URL filtering logs and antivirus reports. Although the focus of such research is not directly related to human subjects, its high fidelity can be a cause of potential harm by exploiting individuals that are part of the dataset.

Monetary Analysis: Measuring the amount of revenue generated by different websites and ad networks is a potential use of such data. Like before, users are not a direct subject of the study, but potentially sensitive human activity provides the basis for researchers to quantify underlying metrics like ad revenue.

5. STAKEHOLDER ANALYSIS

The Menlo Report [6] describes ethical principles for ICT (information and communication technologies) research. Applying those principles to passively collected network data is an important exercise for understanding what research should and should not happen. In this section, we elaborate on some of the specifics of the report along with some data reporting guidelines [1]. These serve as essential considerations when performing research on network tap data (referred in section 6).

Respect For Persons: Individuals who are the subjects under consideration within the tap data should have informed consent of participating in the research as well as the option to opt out. It is the responsibility of the data resource organization to ensure consent. In cases where this is infeasible, researchers handling that data should request for an REB (IRB in the US) waiver and must abide all regulations.

Beneficence: Tap data includes high fidelity personal information of both the individuals and the organization volunteering data. While researchers may explore multiple directions, the essential goal should be the welfare of the community. The overall scope of benefits and risks should be identified to provide a balanced methodology.

Justice: Tap data generally contains a diverse group of subjects. Though this is more relevant for ISP based data, several distinctions may also be made within individuals associated with data obtained from academic institutions and enterprises. All research should be performed in an unbiased manner. In cases where specific cliques are under consideration, the rationale should be explicitly mentioned.

Respect For Law and Public Interest: Research individuals should fully comply with any legal regulations and agreements that might be a part of their research. In the context of network taps, as data is of high fidelity with various venues of exploration, it is incumbent upon researchers not to use the data as a private resource for further analysis.

Data Reporting: Allman and Paxson [1] also provide general guidelines for reporting shared measurement data. Results generated from network tap data should be in an aggregated format to reduce the sensitivity while elucidating a specific trend. Another important aspect is the anonymization of the dataset to rule out

"any" possible affiliation with the providing entity as a means to minimize risk. The next section describes how these policies and regulations fit into research scenarios dealing with network taps.

6. PUTTING IT ALL TOGETHER: INSIGHT INTO NETWORK TAPS

With the ability to record every bit of communication on a link, network taps are a great asset to various organizations. Specific roles within an organization have access to the data collections for administrative and regulatory purposes. Providing similar access of these datasets to researchers without any prior anonymization raises concerns of risk to both the user subjects and the organization itself. In this section, we elaborate on the analysis opportunities from section 4 *specifically* related to tap data, and provide proof of concept examples along with discussions relevant to the ethical considerations at stake. These examples are not an exhaustive list, but rather provide a set from which to generalize how ethics play a fundamental role in measurement research.

Example: A researcher aiming to quantify social network access trends in a university dataset and looking into fluctuations in the performance of overall access times as a result of requests to social networks.

Analysis Techniques: Trend and Performance

Discussion: This research problem provides beneficence to the organization by addressing an important concern. It focuses on evaluating whether social network access substantially affects the performance of other university traffic during work hours. As this is a large scale study, indirectly related to users, it does not pose any major risk to subjects. However, when publishing, informed consent or IRB approval in necessary. In case of reporting results, it is imperative for researchers to remove the specific details for both the subjects and the organization. Furthermore, the data provided must only be used for the allowed purpose.

Example: A study to predict the gender and relationship status of a student at an institution based on their online shopping patterns.

Analysis Techniques: Online Activity and Behavioral

Discussion: As this form of research is a direct study of human subjects, it is essential for researchers to obtain research ethics committee approval. Apart from the high risk to human subjects by exposure to fine-grained information, the research question is of low beneficence to the community and organization. The question is of the trivial nature and the cost of providing access to highly personal information is unlikely to be justified by the impact of the results.

Example: An analysis of revenue generated by malicious websites accessed from within an enterprise network.

Analysis Techniques: Cybercrime & Monetary

Discussion: This study has high risk but also high potential, as it could yield actionable insights for the enterprise related to its bottom line. The analysis of malicious activity can allow the organization to install defensive mechanisms. However, if proper guidelines are not followed, the study might enable malicious entities to attack and benefit from this scenario. As a part of performing cybercrime research, the stakeholders should be kept up to date with all insights, specifically the ones that might pose serious harm to the enterprise.

7. THE WAY FORWARD

While many significant roadblocks exist when conducting research with network tap data, there are many potentially transformative studies which could improve the experience of those using the network every day. When allowing research on these datasets, we propose the use of a combination of the three techniques below to decide how and when to conduct this sensitive research:

- There should be an explicit mention of benefits and operational feedback to the organization volunteering data access.
- Subjects of research should have the option to opt-in for data collection when they start using a network service.
- Researchers should only have access to anonymized data from employees in operational roles within an organization.

8. REFERENCES

[1] M. Allman and V. Paxon. Issues and Etiquette Concerning Use of Shared Measurement Data. In *Proceedings of the ACM IMC* (2007).

[2] M. Allman, E. Blanton, and W. Eddy. A Scalable System for Sharing Internet Measurements. In *Proceedings of PAM* (2002).

[3] M. Foukarakis, D. Antoniades, M. Polychronakis. Deep Packet Anonymization. In *Proceedings of the EUROSEC* (2009).

[4] National Science Foundation. DatCat. http://imdc.datcat.org/.

[5] US Department of Homeland Security. PREDICT. https://www.predict.org/.

[6] US Department of Homeland Security. The Menlo Report.

[7] V. Paxon. Strategies for Sound Internet Measurement. In *Proceedings of the ACM IMC* (2004).

[8] V. Paxson. Internet Traffic Archive. http://ita.ee.lbl.gov/.