An Internet Data Sharing Framework For Balancing Privacy and Utility

Erin E. Kenneally
CAIDA/UCSD *
erin@caida.org

Kimberly Claffy
CAIDA/UCSD
kc@caida.org

1. INTRODUCTION

We re-visit the common assumption that privacy risks of sharing[1] Internet infrastructure data outweigh the benefits, and suggest that we have a window of opportunity in which to apply methods for undertaking empirical Internet research that can lower privacy risks while achieving research utility. The current default, defensive posture to not share network data derives from the purgatory formed by the gaps in regulation and law, commercial pressures, and evolving considerations of both threat models and ethical behavior. We propose steps for moving the Internet research stakeholder community beyond the relatively siloed, below-the-radar data sharing practices and into a more reputable and pervasive scientific discipline, by self-regulating through a transparent and repeatable sharing framework.

The threat model from *not* data sharing is necessarily vague, as damages resulting from knowledge management deficiencies are beset with causation and correlation challenges. And at a more basic level, we lack a risk profile for our communications fabric, partly as a result of the data dearth. Notably, society has not felt the pain points that normally motivate legislative, judicial or policy change – explicit and immediate "body counts" or billion dollar losses. But we must admit, the policies that have given rise to the Internet's tremendous growth and support for network innovations have also rendered the entire sector opaque, unamenable to objective empirical macroscopic analysis, in ways and for reasons disconcertingly resonant with the U.S. financial sector before its 2008 meltdown. The opaqueness, juxtaposed with this decade's proliferation of Internet security, scalability, sustainability, and stewardship issues, is a cause for concern for the integrity of the infrastructure, as well as for the security and stability of the information economy it supports [9].

Strategies to incentivize sharing by amending or enacting legislation merit consideration, and if the past is any indication, communications legislation will eventually be updated to reflect data sharing challenges [5]. However, regulation, especially in the technology arena, is largely reactive to unanticipated side-effects and damages, rather than proactive, fundamental adjustments to predictable difficulties. The behavioral advertising industry is an instructive example; the FTC deferred to industry self-regulation [11] until repeated failures at market-based management of privacy risk induced government action [10]. Furthermore, the length of the legislative policy cycle, confluence of variables involved in changing law, and unclear influence dynamics are out of sync with the need for immediate solutions that interested stakeholder DS and DPs can execute.

Internet research stakeholders have an opportunity to tip the risk scales in favor of more protected data sharing by proactively implementing appropriate management of privacy risks. We seek to advance this objective by outlining a model – the Privacy-Sensitive Sharing (PS2) framework – that effectively manages privacy risks that have heretofore impeded more than ad hoc or nod-&-a-wink data exchanges. Our model integrates privacy-enhancing technology with a policy framework that applies proven and standard privacy principles and obligations of data seekers and data providers. We evaluate our policies and techniques along two primary criteria: (1) how they address privacy risks; and, (2) how they achieve utility objectives.

2. CHALLENGES AND MOTIVATIONS

Researchers have argued that greater access to real network traffic datasets would "cause a paradigmatic shift in computer security research."[1] While data providers (DP) acknowledge the potential benefits of sharing, they are sufficiently uncertain about the privacy-utility risk that they yield to a normative presumption that the risks outweigh potential rewards.

At present, there is no legal framework that prescribes, explicitly incentivizes, or forbids the sharing of network measurement data. Implicit incentives to share measurement data exist, but implementations have mostly floundered. Data sharing relationships that occur are market-driven or organically developed. Unsurprisingly then, there are no widespread and standard procedures for network measurement data exchange. Inconsistent,

*This project is sponsored by the U.S. Department of Homeland Security (DHS) Science and Technology (S&T) Directorate.

[1] By "sharing" we mean any deliberate exchange, disclosure, or release of lawfully possessed data by a
Data Provider (DP) to one or more Data Seekers (DS).

ad hoc and/or opaque exchange protocols exist, but measuring their effectiveness and benefit is challenging. A formidable consequence is the difficulty of justifying resources funding for research and other collaboration costs that incentivize a sharing regime.

Privacy is hard to quantify, as is the utility of empirical research. Both variables are dynamic and lack normative understanding among domain professionals and the general citizenry. In addition to lacking a common vocabulary, the fields of information privacy and network science both lack structured, uniform means of analysis to parse issues and understand and communicate requirements, and common specific use cases. There is no cost accounting formula for privacy liabilities, nor ROI formula for investment in empirical network traffic research. The conundrum, admittedly not unique to Internet research, is that the risk-averse data provider needs utility demonstrated before data is released, and the researcher needs data to prove utility.

The rational predilection against sharing is strengthened by an uncertain legal regime and the social costs of sensationalism-over-accuracy-driven media accounts in cases of anonymized data being reverse engineered. Although there is interest in efficient sharing of measurement data, it hangs against a backdrop of legal ambiguity and flawed solution models. This backdrop and our experiences with data-sharing inform the Privacy-Sensitive Sharing framework (PS2) we propose.

2.1 An Uncertain Legal Regime

The concept of personally identifiable information (PII) is central to privacy law and data stewardship. Ambiguity over this fundamental concept drives privacy risk assessment. Privacy presumes identity so unless identity is defined in relation to network data artifacts, the notions of privacy and PII are already disjointed in the Internet data realm. Both the legal and Internet research communities acknowledge this unclarity regarding PII in Internet data – its definition is context-dependent, both in terms of technology and topology. Definitional inconsistencies are exacerbated as evolution of technologies and protocols increases capabilities and lowers the cost of linking network data to individuals. Against this dynamic interpretation of PII, many privacy-related laws must find practical application.

Much of the risk management challenge lies in this linguistic incongruity between the legal and technical discourse about traffic data – its definitions, classifications and distinctions. The legal perspective apportions different risk to content versus addressing based on a necessarily limiting analogy to human communications versus machine instructions, respectively. Similarly, officers of the court associate IPAs with a greater privacy risk than URLs, based on our past and still partial ability to link those to an individual. This distinction was always artificial (albeit not unfounded) since both types of data reference a device or virtual location rather than an individual, and many URLs directly reveal much more user information than an IP address.

Our jurisprudence contributes to (or reflects) our cognitive dissonance about privacy. The U.S. legal regime has not consistently or comprehensively protected IPAs or URLs as PII, and caselaw largely fails to recognize a reasonable expectation of privacy in IPAs and URLs (i.e., no Fourth Amendment constitutional protection). Privacy statutes like HIPAA explicitly include IPA as protected personal data, while the majority of state data protection laws do not. Juxtaposed with this disparity is a reasonable if not normative societal expectation of privacy in those traffic components (i.e, IPA or URLs with search terms) as the digital fingerprints between anonymous bits and their carbon-based source. Technologies developed and deployed to strengthen the privacy of these components, e.g., IPA masking services such as Tor, in-browser automated URL deletion features, suggest their privacy is considered a social good.

And yet, IPA data is the principal evidentiary underpinning in affidavits for search warrants or subpoenas relied upon by private intellectual property owners (e.g. RIAA) and government investigators to identify and take legal action against the person associated with the IPA. In practice there is little functional differentiation between IPAs/URLs and other, privacy-protected PII, yet the related legal treatment is far less consistent.

2.2 Flawed Technology Models

In addition to muddy legal waters, technology models for data-sharing proposed thus far have failed to accomplish their own objectives. Efforts by the technical community focus on improving computing technologies to solve the privacy problem, in particular anonymizing data components considered privacy-sensitive. Since privacy risk is influenced by shifting contexts associated with relationships between people, data, technology and institutions, these "point" solutions inherently fail to address the range of risks associated with the lifecycle of shared data. Like networked systems themselves, the privacy threat is evolving, unpredictable, and in constant tension with user needs. It must therefore be managed accordingly, including creating and reinforcing conditions that allow privacy technology to be effective.

Researchers have advocated and supported sharing data for years [2, 16, 7]. Recognizing that these purely technical efforts have had limited success in supporting needed cybersecurity research, DHS developed the PREDICT project to formalize a policy framework for balancing risk and benefit between data providers and researchers. A policy control framework enables the technical dials to allow more privacy risk if a specific use justifies it. A classic example is how to anonymize IPAs critical to answering research questions – traces protected by prefix-preserving anonymization may be subject to re-identification risk or content observation risk, but policy controls can help data providers minimize the chances that sensitive information is misused or wrongfully disclosed.

2.3 Privacy Risks of Internet Research – the Court of Public Opinion

The privacy risks of data sharing fall into two categories: disclosure and misuse. Although we focus on the risks of data sharing, our model can also be applied to initial data collection, which also poses significant privacy risks that should not be ignored.

Public disclosure is the act of making information or data readily accessible and available to the general public via publication or posting on the web, including log files with IP addresses of likely infected hosts.

Accidental or malicious disclosure is the act of making information or data available to a third party(s) as a result of inadequate data protection [4].

Compelled disclosure to third parties risk arises with the obligations attendant to possessing data, such as having to respond to RIAA subpoenas requesting data disclosure in lawsuits.

Government disclosure involves the release of data to government entities, illustrated in the NSA wiretapping revelations and subsequent EFF lawsuits [8].

Misuse of user or network profiles such as with network traffic that contains proprietary or security-sensitive information, or reveals user behaviors, associations, preferences or interests, knowledge that attackers – or advertisers – can then exploit.

Inference misuse risk involves synthesizing first- or second-order PII to draw (possibly false and damaging) implications about a persons behavior or identity.

Re-identification or de-anonymizing misuse risk. Anonymization involves obfuscating sensitive PII by replacing it completely or partially with synthetic identifiers, or using aggregation or statistical techniques intended to break the connection between the persons and reference identifiers. Re-identification or de-anonymization, conversely, involves reversing data masks to link an obfuscated identifier with its associated person. Shared anonymized data poses a misuse risk because it is vulnerable to reidentification attacks that make use of increasingly available public or private information beyond the knowledge or control of the original or intermediate data provider [15, 13].

De-anonymization risk bears special consideration in the growing incongruity around PII. DPs face increasing legal and societal pressures to protect the expanding amounts of PII they amass for legitimate business purposes. Yet, DPs are under equal pressure from the marketplace to uncover and exploit PII in order to better connect supply and demand and increase profit margins on their goods and services. DPs will have a growing interest in anonymization to avoid triggering laws that hinge on definitions of PII which exempt aggregate or anonymized data. In the meantime, both legitimate and criminal consumers of PII are motivated to advance de-anonymization techniques to uncover sensitive identity data, as well as strategies to extract informatiotn without direct access to data by 'sending code to the data' [12]. Similar to the arms race between exploits and defenses in the systems security arena, de-anonymization techniques will become commoditized, lowering the barrier to extracting PII for investigative reporting, law enforcement, business intelligence, research, legal dispute resolution, and the presumed criminal threatscape.

2.4 Communicating the Benefits of Research

Internet measurement supports empirical network science [6], which pursues increased understanding of the structure and function of critical Internet infrastructure, including topology, traffic, routing, workload, performance, and threats and vulnerabilities. But network researchers and funding agencies have yet to outline a network science agenda [3], partly due to their lack of visibility into the nation's most critical network infrastructure, but also because the field is quite young relative to traditional scientific disciplines. In light of this challenge, we offer the following criteria to help measure and communicate empirical network research utility:

- Is the objective for data use positively related to social welfare?
- Is there a need for such empirical research?
- Is such network research already conducted?
- Could the research be conducted without the data?
- Is sufficiently similar data already being collected elsewhere that could be shared?
- Are research and peer review methods as transparent, objective, and scientific as possible while being responsible to privacy concerns?
- Can the results be acted upon meaningfully?
- Are the results capable of being integrated into operational or business processes? Or security improvements, e.g., situational awareness of critical infrastructure?

3. PS2 FRAMEWORK: DESCRIPTION AND EVALUATION

We propose a repeatable process for data sharing which employs and enforces these techniques – a process that supports sharing of not only the data, but also the obligations and responsibilities for privacy risk. We describe our Privacy-Sensitive Sharing Framework and then evaluate its ability to address the privacy risks outlined in 2.3 and the utility criteria in 2.4. A core principle of the framework is that privacy risks associated with shared Internet data are contagious – if the data is transferred, responsibility for containing the risk lies with both provider and seeker of data. Recognizing that privacy risk management is a collective action problem, the PS2 hybrid framework contains this risk by replicating the collection, use, disclosure and disposition controls over to the data seeker.

Our strategy presumes that the DP has a lawful ownership or stewardship right to share the network data, and that the associated data privacy obligations travel commensurate with those rights. Sharing data does not discharge the DP from control or ownership obligations. This inextricable connection is essential to engendering the trust that facilitates data exchanges.

3.1 Components of the PS2 Framework

The components of our framework are rooted in the principles and practices that underlie privacy laws and policies on both the national and global levels.[2]

- Authorization – Internal authorization to share with exchange partner(s) involves explicit consent of the DP and DS, and may require the consent of individuals identified or identifiable in network traffic, which can often be implicit by way of proxy consent with the DP. Requirements for consent to Internet traffic monitoring are unresolved, but will no doubt be a part of forthcoming legal, policy and community decisions. In addition, the DP and DS should obtain some external explicit authorization to share data, such as through an external advisory or publication review committee, or an organization's Institutional Review Board (IRB).
- Transparency – The DP and DS should agree on the objectives and obligations associated with shared data. Terms might require that algorithms be made public but data or conclusions protected, or vice-versa. Transparency may require validating analytic models or algorithms. It should be possible to apply scientific principles of testability and falsifiability to analytical conclusions.
- Compliance with applicable law(s) – Collection and use of data should comport to laws that have an authoritative interpretation about proscribed behaviors or mandated obligations.
- Purpose adherence – The data should be used to achieve its documented purpose for being shared.
- Access limitations – The shared data should be restricted from non-DS partners (government, third parties) and those within the DS organization who do not have a need and right to access it.
- Use specification and limitation – Unless otherwise agreed, the DP should disallow merging or linking identifiable data contained in the traffic data.
- Collection and Disclosure Minimization – The DS should apply privacy-sensitive techniques to stewardship of the network traffic such as:

  A. Deleting all sensitive data.
  B. Deleting part(s) of the sensitive data.
  C. Anonymizing, or otherwise de-identifying all or parts of sensitive data.
  D. Aggregation or sampling techniques, such as scanning portions of networks and generalizing results.
  E. Mediation analysis, e.g., 'sending code to the data' (or a person) in lieu of sharing data.
  F. Aging the data such that it is no longer current or linkable to a person.
  G. Size limitations, e.g., minimizing the quantity of traces that are shared.
  H. Layered anonymization, e.g., applying multiple levels of anonymization are applied to lower the de-identification countermeasure effectiveness.

- Audit tools – Techniques for provable compliance with policies for data use and disclosure, e.g., secure audit logging via a tamper-resistant, cryptographically protected device connected to but separate from the protected data, accounting policies to enforce access rules on protected data.
- Redress mechanisms – Technology and procedures to address harms from erroneous use or disclosure of data, including feedback mechanism to support corrections of data or erroneous conclusions.
- Oversight – Following authorization, any systemic sharing or program needs subsequent third-party checks and balances such as Institutional Review Boards (IRB), external advisory committees, or a sponsoring organization.
- Quality data and analyses assurances – Processes should reflect awareness by DS and DP of false positives and inference confidence levels associated with the analyses of shared, privacy-sensitive data.
- Security – Controls should reasonably ensure that sensitive PII is protected from unauthorized collection, use, disclosure, and destruction.
- Training – Education and awareness of the privacy controls and principles by those who are authorized to engage the data.
- Impact assessment – Experiment design should consider affected parties and potential collateral effects with the minimal standard of doing no further harm [14]. Considerations include possible psychological harm, physical harm, legal harm, social harm and economic harm.

[2] In particular, the Fair Information Practices (FIPS) are considered de facto international standards for information privacy and address collection, maintenance, use, disclosure, and processing of personal information.

- Transfer to third parties – Further disclosure should be prohibited unless the same data control obligations are transferred to anyone with whom the data is shared, proportionate to the disclosure risk of that data.
- Privacy Laws – Existing laws can provide enforceable controls and protections that can help shape sharing dynamics. For example, the Electronic Communications and Privacy Act has well-defined parameters for traffic content disclosure based on whether the DS is a government entity.

3.2 Vehicles for Implementing PS2s

Given the legal grey areas and ethical ambiguity around disclosure and use of network measurement data discussed in Section 2.1, we recommend MOUs, MOAs, model contracts, and binding organizational policy as enforceable vehicles for addressing privacy risk both proactively and reactively. For intended wider disclosure of data it may be cost-preferential and risk proportional to release data under a unidirectional (blanket) AUP, which trades off lower per-sharing negotiation costs associated with bilateral agreements for potentially lower DS privacy protections and greater DP enforcement and compliance costs.

The research community would benefit from developing and publishing a reference sets of MOUs which embed this or a similar framework, in support of a common interpretation of acceptable sharing and reasonable practices. Recognizing that the law is not a one-way proscription, but ultimately reflects and institutionalizes community norms, network researchers have a window of opportunity to pre-empt privacy risks in legally grey areas, while informing policymakers of possibilities they are not well-positioned to discern themselves.

3.3 How well the PS2 addresses privacy risks

We evaluate the effectiveness of the Privacy Sensitive Sharing Framework in balancing the risks and benefits by: (1) assessing how it addresses the privacy risks outlined in 2.3; and (2) assessing whether it impedes the utility goals described in 2.4. Tables 1 and 2 illustrate the value of the hybrid policy-technology model.

Although the risks and benefits we enumerate are not all-inclusive, they are a strong foundation for engaging in privacy-utility risk management. For DS, it provides a tool to examine whether the proposed research balances the privacy risks and utility rewards. For an oversight committee, it helps determine whether possible risks are justified, by explicitly asking the user to assess sharing risks against technical and policy controls, as well as to assess the achievement of utility goals against those controls. For prospective DP, the assessment will assist the determination of whether or not to participate. This framework can also be applied as a self-assessment tool for data sharing engagements.

3.4 How well the PS2 promotes utility goals

Table 1 suggests that the Minimization techniques (the technical controls discussed in 3.1) partly or completely address privacy risks, implying little need for a policy control backdrop. However, when evaluating how the Minimization techniques fare against the Utility Goals in Table 2, they mostly fail, revealing the limits of a one-dimensional technical approach. This failure is unsurprising, since data minimization techniques intentionally obfuscate information often essential to network management and troubleshooting, countering security threats, and evaluating algorithms, applications, and architectures. A purely technical approach breaks down along the utility dimension, justifying a hybrid strategy that addresses both privacy risk and utility goals.

An evaluation of the PS2 framework should also consider practical issues such as education costs, whether new privacy risk(s) are introduced, whether control(s) are forward-looking only or address legacy privacy risks, and free rider problems created by NPs who choose not to share.

4. CONCLUSIONS

We proposed a Privacy Sensitive Sharing framework that offers a consistent, transparent and replicable evaluation methodology for risk-benefit evaluation rather than relying on subjective, opaque and inconsistent evaluations that turn on *"trust me"* decision metrics. In designing the framework we have considered practical challenges confronting security professionals, network analysts, systems administrators, researchers, and related legal advisors. We have also emphasized the proposition that privacy problems are exacerbated by a shortage of transparency surrounding the who, what, when, where, how and why of information sharing that carries privacy risks. Transparency is defeated if one of the sharing parties is dishonest about its objectives, or if the DP or DS's organizational motivations supersede the desire to engender trust. We developed PS2 to enable transparency as a touchstone of data-sharing.

Documented experimentation with protected data-sharing models like PS2 will allow those professionals with access to ground truth about privacy risks and measurement rewards to practically influence policy and law at these crossroads. Indeed, we believe a window of opportunity exists in this time and space of uncertainty in interpreting and applying privacy-related laws to the sharing of network traffic and other data for research. Rather than wait for law and policymakers to magically "get it right", it behooves infrastructure operators and researchers to self-regulate, until and unless and in the interest of informing clarity from top down regimes. Information security controls were initially considered a financial liability until regulations rendered lack of security a compliance liability. We optimistically anticipate circumstances that reveal that rather than data-sharing being a risk, *not* sharing data is a liability. We offer the PS2 as a tool to help community mindset(s) move (or in some cases, already moving) in that direction.

Privacy Risk/PS2	Public Disclosure	Compelled Disclosure	Malicious Disclosure	Government Disclosure	Misuse	Inference Risk	Re-ID Risk
Authorization		X	X		X	X	X
Transparency	X	X	X	X	X		
Law Compliance			X			X	X
Access Limitation		X			X	X	X
Use Specification		X	X		X	X	
Minimization							X
Audit Tools	X	X	X	X	X	X	X
Redress	X	X	X	X	X	X	X
Oversight		X	X			X	X
Data Quality	X	X	X	X			X
Security		X				X	X
Training/Education		X	X			X	X
Impact Assessment	X	X	X	X	X		

Table 1: Privacy risks evaluated against the PS2 privacy protection components. (*Minimization* refers to the techniques evaluated in Table 2.)

Min. Tech. / Utility	Is Purpose Worthwhile?	Is there a need?	Is it already being done?	Are there alternatives?	Is there a scientific basis?	Can results be acted upon?	Can DS & DP implement?	Reasonable education costs?	Forward & backward controls?	No new privacy risks created?	No free rider problem created?
Not Sharing	X	X	X	X	X	X	X				
Delete All	X	X	X	X	X	X	X		X		
Delete Part	X	X		X	X		X		X	X	
Anonymize	X	X	X	X	X		X	X	X	X	
Aggregate	X	X	X	X	X				X	X	
Mediate (SC2D)	X						X	X			X
Age Data	X	X	X	X	X		X			X	
Limit Quantity	X	X	X	X	X	X	X		X	X	
Layer Anonymization	X	X	X		X	X	X	X	X		

Table 2: PS2 minimization (of collection and disclosure) techniques evaluated against utility.

5. REFERENCES

[1] Allman, M., and Paxson, V. Issues and etiquette concerning use of shared measurement data. In *IMC* (2007).

[2] Allman, M., Paxson, V., and Henderson, T. Sharing is caring: so where are your data? *ACM SIGCOMM Computer Communication Review 38*, 1 (Jan 2008).

[3] Assocation, C. R. Network Science and Engineering (NetSE) Council, 2008. http://www.cra.org/ccc/netse.php.

[4] Barbaro, M., and T. Zeller, J. A Face is Exposed for AOL Searcher No. 4417749. *New York Times* (Aug 2006).

[5] Burstein, A. Amending the ecpa to enable a culture of cybersecurity research. *Harvard Journal of Law & Technology 22*, 1 (December 2008), 167–222.

[6] C. B. Duke, *et al.*, Ed. *Network Science*. The National Academies Press, Washington, 2006.

[7] CAIDA. DatCat. http://www.datcat.org.

[8] Cauley, L. NSA has massive database of Americans' phone calls. *USA Today* (May 2006).

[9] Claffy, K. Ten Things Lawyers should know about Internet research, August 2008. http://www.caida.org/publications/papers/2008/.

[10] Commission, F. T. FTC Staff Report: Self-Regulatory Principles For Online Behavioral Advertising, February 2009. http://ftc.gov/os/2009/02/P085400behavadreport.pdf.

[11] Federal Trade Commission. Online behavioral advertising: Moving the discussion forward to possible self-regulatory principles, 2007. http://www.ftc.gov/os/2007/12/P859900stmt.pdf.

[12] Mogul, J. C., and Arlitt, M. Sc2d: an alternative to trace anonymization. In *SIGCOMM MineNet workshop* (2006).

[13] Narayanan, A., and Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on Security and Privacy* (2008). http://www.cs.utexas.edu/ shmat/shmat_oak08netflix.pdf.

[14] "NIH". The Belmont Report - Ethical Principles and Guidelines for the protection of human subjects of research.

[15] Porter, C. De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information. *Shilder Journal of Law, Communication, and Technology*, 3 (Sept. 2008). http://www.lctjournal.washington.edu/Vol5/a03Porter.html.

[16] Vern Paxson. The Internet Traffic Archive. http://ita.ee.lbl.gov/.