The Cooperative Association for Internet Data Analysis

www.caida.org > data : : sharing

Promotion of Data Sharing

Internet research relies on a wide variety of data on the structure, dynamics, and usage patterns of operational Internet infrastructure, for parameterization and validation of scientific modeling and analysis efforts. As our use of and dependence on the Internet expands, an expanding range of data of interest and utility to an increasing number of disciplines must bring with it deeper consideration of privacy in collaboration and data sharing models, especially between industry and academia.

We have proposed to move the Internet research stakeholder community beyond the relatively siloed data sharing practices and into a more reputable and pervasive scientific discipline, by self-regulating through a transparent and repeatable sharing framework. Our model -- the Privacy-Sensitive Sharing (PS2) framework -- integrates privacy-enhancing technologies with a policy framework that applies proven and standard privacy principles and obligations of data seekers and data providers, in coordination with techniques that implement and provably enforce those obligations. The PS2 framework considers practical challenges confronting security professionals, network analysts, systems administrators, researchers, and legal advisors. It embodies the proposition that privacy problems are exacerbated by a shortage of transparency surrounding the who, what, when, where, how and why of sharing privacy-sensitive information We evaluate our framework along two primary criteria: (1) how they well the policies and techniques address privacy risks; and, (2) how well policies and techniques achieve utility objectives. Below we excerpt from this paper, including a review of the practical risks and benefits of data sharing, as well as the motivation, components, and evaluation of our model.

Information on CAIDA's datasets can be found in the overview of CAIDA datasets page. Most of CAIDA's datasets are also indexed in DatCat, the Internet Measurement Data Catalog.

1. Motivation

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. The threat model from not data sharing is necessarily vague, as damages resulting from knowledge management deficiencies are beset with causation and correlation challenges. More fundamentally, we lack a risk profile for our communications fabric, partly as a result of the data-sharing 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. Admittedly, 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 the information economy it supports. [10].

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. Transparent and morally defensible self-regulation in the interests of building social capital and informing legal and judicial regimes will allow stakeholders to more practically influence policy and law at these crossroads. Information security controls were initially considered a liability (from a cost perspective) until regulations rendered lack of security a compliance liability. We anticipate circumstances to reveal that rather than data-sharing being a risk, *not* sharing data is a liability. We offer the PS2 as a tool to help move the community mindset in that direction as productively and safely as possible.

2. Challenges

The strategic challenge is similar to other domains: how to balance utility goals with privacy risks for data seeker (DS) and data providers (DP). Internet researchers and systems security personnel are generally DS - entities seeking to share, responsibly disclose, acquire or otherwise exchange lawfully possessed network data. 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. 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, 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 for research and other collaboration costs that incentivize a sharing regime. On the other hand, the high cost of independently acquiring datasets is a motivation for re-use where possible.

Privacy is difficult to quantify, as is the utility of measurement-based research. Both variables are dynamic and lack normative understanding among both domain professionals and the general citizenry. As fields of study, privacy and network science are both hindered by the absence of: common vocabulary, open and configurable reference models, uniform means of analysis, common sets of use cases, and unsurprisingly, any standard cost (liability) accounting or ROI formulas. A circular conundrum 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. While there are no procedures or regulatory framework to foster widespread exchange, there is also no framework that prohibits it either. Although there is interest in efficient and widespread 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 our privacy-sensitive sharing framework (PS2).

2.1 An Uncertain Legal Regime

At least in the U.S. and European Union (EU) regulatory regimes, the concept of personally identifiable information (PII) is central to privacy law and data stewardship in general. Unlike the EU model which allots overarching protection for PII, the U.S. protects PII across a patchwork of caselaw, state and federal industry-specific laws covering health data, financial data, education data, employment data, insurance records, government-issued records, credit information, and cable and telephone records. Although the definition of PII bears common threads across sectors, it is nonetheless fractured along a continuum of *first* and *second-order* identifiers (defined in Section 3.1). Crafting frameworks that generalize PII across domains, and or support cross-domain information necessitates attaching to the most expansive definition.

At its core, there is ambiguity over fundamental concepts upon which privacy risk assessment turns. First, 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 that the concept of PII in Internet data is not clear - its definition is context-dependent, both in terms of technology and topology. Further, the ability to link network data to individuals - as well as the cost of doing so - changes over time as technologies and protocols evolve. Yet, PII is fundamental to interpreting and applying many laws. Most notably is the United States' primary law covering the privacy of network traffic: the Electronic Communications and Privacy Act (ECPA), which provides statutory privacy protection for the interception and disclosure of certain electronic communications.

For example, blanket characterizations of IPA or URLs as PII (or not) are necessarily inaccurate because they alone cannot capture the range of privacy risks - either category could include an instance with PII, but most observed instances do not. In practice there is little functional differentiation between these traffic components and other, privacy-protected PII, yet the related legal treatment of IPAs and URLs is far less consistent. A more accurate risk assessment depends on knowing who collected it and how they use, disclose, and dispose of the traffic data.

The risk management challenge lies in the linguistic incongruity between the legal and technical discourse about traffic data - its definitions, semantic classifications and interpretations. Officers of the court associate IPAs with a greater privacy risk than URLs based on our past and still partial ability to link IPAs to an individual. This distinction was always artificial (albeit not totally 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.

More specifically, this legal-technical gap exposes privacy risks with network operational data insofar as many laws do not explicitly allow for research use of network data [4], and there is no bright line caselaw applying their respective exceptions to the context of sharing Internet data for research.

2.2 Flawed Technology Models

Most data-sharing efforts by the networking research community focus on improving computing technologies to solve the privacy problem, with anonymization commanding the bulk of the attention. A typical researcher approach is to enumerate the possibly privacy-sensitive information present in network traffic traces, and then implement a technical, typically cryptographic, solution to replace this information completely or partially with synthetic identifiers, normally implemented by encrypting or otherwise removing all or part of identifiers.

Since privacy risk is influenced by evolving contexts associated with relationships between people, data, technology and institutions, solely technical solutions are inherently insufficient to balance the privacy/utility tradeoff. Technical researchers may rightly ask why we would predicate sharing architectures on ambiguous, unquantifiable and fallible human trust enforced by law and policy, if we can build trust through technology. The response is simple: while a purely technical approach may significantly ameliorate privacy risk, it largely fails to render empirically grounded answers to most questions being asked about the Internet today.

For example, while anonymization schemes can enhance the privacy of IPA in shared network traces, if it removes the

ability to do geographic or topological analysis, the research utility of that data for studying DDoS modus operandi is dramatically reduced. A policy control framework enables the technical dials to allow for more privacy risk if a specific use justifies it. For example, 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 Reactive Top-Down Policy

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 the evolved needs from last few decades. However, regulation, especially in the technology arena, is largely reactive to unanticipated side-effects and dangers rather than making proactive, fundamental adjustments to predictable difficulties. Further, the length of the legislative policy cycle, confluence of variables involved in changing law, and unpredictable change agents are not amenable to immediate solutions that interested stakeholder DS and DPs can execute. In the Internet measurement space, a legislative solution means awaiting the familiar change agent aforementioned: body counts or billion dollar losses that result from the lack of ground truth about the structure and function of networks that comprise our critical communications infrastructure.

3 Sharing risks and benefits

3.1 Who, What, When

Who is at risk when network data is shared?
Entities potentially at risk when network traffic is shared include: persons who are identified or identifiable in network traffic, researchers, and network providers (NP) such as ISPs, backbone providers, and private network owners. In addition to legal liabilities and ethical responsibilities, researchers and their institutions also risk withdrawal of data and or funding as a result of privacy leakage. Society also bears costs associated with misinformation, mistrust, and internalizing behavioral norms that may result from privacy harms.

Which traffic data components are privacy-relevant?
We call a *first-order identifier* one which functionally distinguishes an individual: first and last name, social security number, government-issued and other account identifiers, physical and email addresses, certain biometric markers, and possibly the same information about immediate family. A *second-order identifier* could be an IP address (IPA), machine access code (MAC) address, host name, birthdate, phone number, zip code, gender, and financial, health, or geographic information. These indirect identifiers can also include aggregated or behavioral profile information such as IP header information, which in many cases can reveal which applications are used, how often, and with which machines. Indirect identifiers also include URL click streams, which can reveal information about the content of communications, including search terms.

Under what conditions do these data types pose risk?
Network traffic measurement data can present a privacy risk when information in packets and flow records can directly expose non-public information about persons - such as health, sexual orientation, political affiliation, religious affiliation, criminal activity, associations, behavioral activities, physical or virtual location; or, organizations - such as intellectual property, trade secrets or other proprietary information. Network traffic may also indirectly expose non-public, sensitive information if correlated (linked) with other public or private data, such as the case with lists of IPAs of worm-infected and thus vulnerable hosts. Network data can also yield mistaken attributions and inferences about behavior, potentially more damaging than correct inferences.

The privacy risk across time may also vary, as the threat may be immediately manifest upon disclosure of data, or it may be a latent risk which is held in abeyance until some future condition arises. Lack of transparency between the DP and DS regarding the shared data's nature, scope, and lineage is invariably a condition that enhances risk.

3.2 Privacy Risks of Internet Research - Laws and Courts of Public Opinion

It is impractical to enumerate all laws that may affect privacy risk, but such an enumeration is not prerequisite to capturing the foreseeable risks of network data sharing. It is sufficient to note that legal liability or ethical obligations underlie each privacy risk. Dismissing ethical obligations as discretionary and unenforceable overlooks how ethical violations are treated by public opinion, and also ignores the fact that many laws are an evolution of ethical norms. In the U.S., privacy-related legal liabilities can derive from the federal Constitution (most notably the Fourth Amendment), federal law and regulation, contract law, tort law (e.g., invasion of privacy), state law equivalents, and organizations' privacy policies. Beyond the legal risk, violations of ethical obligations can create normative harms that implicate reputation and financial damages.

We break down the privacy risks of data sharing into two categories: disclosure and misuse.

Public disclosure is the act of making information or data readily available to the general public via publication or posting on the web. The privacy risks of sharing data containing PII which is subsequently displayed on the web are obvious and incontrovertible. More common and challenging are publicly available network traces and activity logs which reveal

identifying information about infected hosts. Such disclosure raises the risk that unpatched or vulnerable hosts will be further exploited, thus creating security and reputation risks for individuals and organizations.

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. AOL provided a quintessential example in 2006 when they released an anonymized data set of search queries that, with sufficient public meta-data were linked back to users conducting the searches who were then exposed in the NYT. [2]

Compelled disclosure to third parties risk arises with the obligations attendant to possessing data, such as having to respond to subpoenas requesting data disclosure in lawsuits. The RIAA campaign to massively subpoena ISPs and universities in an attempt to identify copyright infringers is a notorious example. To illustrate, many entities (including research organizations) have chosen not to retain traffic statistics of operational and research interest, to avoid any such compulsion.

Government disclosure involves the release of data to government entities. An infamous example is the disclosure of call data records by major telecommunications carriers to the National Security Agency around 2007 [6]. Release to the government introduces another level of risk involving civil rights and liberties, such as imprisonment and restrictions on speech and associations.

Misuse of user or network profiles arises with network traffic that contains information about proprietary or security-sensitive network architectures or business operations. Advancing traffic and topology analysis, data mining and classification techniques can derive sensitive information from seemingly benign traffic data, and thereby reveal user behaviors, associations, preferences or interests, which attackers, advertisers, or content owners can then exploit. Network operators themselves may use such information for network management, illustrated by Comcast's recent throttling of BitTorrent traffic. The relatively invasive traffic engineering technique, combined with a lack of transparency in deploying it, led to public uproar and an unprecedented FCC regulatory ruling.

Inference misuse risk involves synthesizing first-order or second-order identifiers to draw inferences about a person's behavior or identity.

Re-identification and De-anonymizing misuse involves reversing data anonymization or masks to link an obfuscated identifier with its associated person. Shared anonymized data poses a misuse risk because it is variably vulnerable to re-identification attacks using public or private information whose (increasing) availability is beyond the knowledge or control of the original or intermediate data provider [14]. Anonymized data may not immediately expose PII, but any time a piece of de-identified data has been linked to first order identifying information, other anonymous aspects of the obfuscated data are easier to de-anonymize. Aggregation or statistical techniques for anonymization are not immune to re-identification risk. Examples of reidentification risk are the 2007 Netflix prize incident [12], and a similarly embarrassing episode of re-identification within the Internet research community [1].

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 mine PII in order to better connect supply and demand, and increase profit margins on their goods and services. DPs will turn to anonymization to avoid triggering privacy laws that exempt aggregate or anonymized data.

Like the arms race between exploits and defenses in the systems security arena, de-anonymization techniques will likely become commoditized to support investigative reporting, law enforcement, business intelligence, research, legal dispute resolution, and the presumed criminal threatscape. Several state legislatures have enacted laws to ban the release of sensitive private information because of this re-identification risk [7], although these are not contention-free. Re-identifications concerns motivated the National Human Genome Research Institute's recent removal of open access to the pooled genomics data it posted on the Internet in 2006 [9].

3.3 Utility of Internet Measurement

The benefits of network traffic measurement derive from the value of empirical network science [5], which includes a better understanding of the structure and functions of networks that comprise critical Internet infrastructure. Network researchers and funding agencies struggle to establish a science agenda, partly due to their lack of visibility into the infrastructure, but also because the field is younger and less well-defined than traditional scientific disciplines.

The following criteria help measure and communicate empirical network research utility:

- The objective for sharing the data produces or promotes social welfare or generalizable knowledge.
- The network research data is not already being shared, or if it is, there remains a qualitative need for sharing between other DS and DPs
- The research could not be conducted without the shared data

- The scientific methodology using the shared data is transparent, objective, and repeatable relative to any privacy controls that are implemented.
- Research results can be acted upon meaningfully.
- Research results can be integrated with business processes or security operations, such as situational awareness of critical infrastructure.

Research that could satisfy the above criteria include:

- information and network security questions regarding system threats, including characterizing baseline and anomalous workloads, modeling malware, developing effective strategies to deal with threats.
- macroscopic analysis of Internet topology; understanding the how the evolution of the network is affecting the efficiency and capabilities of the underlying routing, transport, and naming protocols.
- understanding the effect of the prevalence and growth of new applications on Internet workload, topology, and infrastructure economics.
- validation of traffic, congestion control, and performance assumptions, models, and analyses, both for current and proposed new technologies.
- development and evaluation of new technology, including measurement and sampling techniques.

4 PS2 Framework: Elements, Execution, and Evaluation

We describe the Privacy-Sensitive Sharing Framework and then evaluate the model's ability to address the privacy risks outlined in 3.2 and the utility criteria in 3.3. Recognizing that privacy risk management is a collective action problem, our PS2 framework contains this risk by replicating the collection, use, disclosure and disposition controls over to the DS. This framework contemplates that the privacy risks associated with shared data are contagious - if the data is transferred, responsibility for containing the risk lies with both provider and seeker of data. In other words, there is no automatic detachment of control or ownership by the DP when the data is shared.

4.1 Elements of PS2

While not framed around specific legislation, The components of our framework are rooted in principles and practices that underlie privacy laws and policies on both the national and global levels. 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. The FIPs have spawned a series of authoritative reports, guidelines, and model codes that implement these principles [13]. The PS2 is an attempt to apply these principles to the context of Internet measurement and sharing, aiming to build a touchstone for ethically defensible sharing scenarios.

- **Authorization** - Internal authorization to share requires explicit consent of the DP and DS, and may require consent of individuals identifiable in network traffic, which can often be implicit via proxy consent with the DP.
- **Oversight** - The DP and DS should obtain some external oversight of the proposed sharing, such an Institutional Review Boards (IRB).
- **Transparency** - The DP and DS should agree on the objectives and obligations associated with shared data. Data-sharing terms might require that the algorithms be public but that the data and or conclusions remain protected, or vice versa [15].
- **Compliance with applicable law(s)** - Collection and use of data should comport to a reasonable if not case-law precedented interpretation of laws that speak directly and clearly to sharing risks about proscribed behaviors or mandated obligations.
- **Purpose adherence** - The data should be used to try to achieve the documented goal of sharing.
- **Access limitations** - The shared data should be restricted from those who do not have a need and right to access the shared data.
- **Use specification and limitation** - Unless otherwise agreed, The DP should deny 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:
  1. Deleting sensitive data.
  2. Deleting part(s) of the sensitive data.
  3. Anonymizing Hashing De-Identifying all or parts of the sensitive data.
  4. Aggregating or sampling.
  5. Mediation analysis Human Proxy - a sandbox approach that involves "sending the code to the data" rather than releasing sensitive data for analyses.
  6. Aging the data - such that traffic that contains sensitive data that is non-current, i.e., no longer a direct or indirect identifier.
  7. Size quantity limitation - minimizing the quantity of traces shared.
  8. Multiple layers of anonymization.
- **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** - Procedures to address harms from inappropriate data use or disclosure, including a feedback mechanism to support correction of datasets and or erroneous conclusions.
- **Quality data and analyses assurances** - Awareness by the DS and DP of inference confidence levels associated with the data.
- **Security** - Controls should reasonably ensure that sensitive PII is protected from unauthorized collection, use, disclosure, and destruction.
- **Training** - some level of education and awareness of the privacy controls and principles by those who are authorized to engage the data.
- **Impact assessment** - Research design should consider potential collateral effects on affected parties, and seeks methods that do no further harm.
- **Transfer to third parties** - prohibited unless the same data control obligations are transferred, relative to the disclosure risks associated with that data.

4.2 Execution of PS2

To navigate the legal and ethical ambiguity around disclosure and use of network measurement data discussed in Section 2.1, we propose Memoranda of Understanding (MOUs), Memoranda of Agreement (MOAs), model contracts, and binding organizational policy as enforceable vehicles for addressing privacy risk both proactively and reactively. For less privacy-sensitive data, a unidirectional Acceptable Use Policy AUP may be cost-preferential to negotiating bilateral agreements. Explicit consent about controls for shared data provides an enforceable standard and certainty that can serve as a safe harbor for liability under data privacy laws.

4.3 Evaluation of PS2

The PS2 framework facilitates a rigorous examination of whether the proposed research balances privacy risks and utility rewards. For an oversight committee, it helps determine whether possible risks are justified, by specifically 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 the prospective DP, the assessment will assist the determination whether or not to participate.

Table 1 assesses whether the privacy risks are mitigated by the primary components of PS2. The X's indicate that the particular PS2 policy component in the row fails to mitigate against the privacy risk enumerated in the corresponding column. The table starkly shows that purely policy components of PS2 still leave wide gaps in addressing the full range of privacy risks. Further, it suggests that technical minimization techniques (Section 4.1) can address all privacy risks, implying the sufficiency of a purely technical sharing framework in lieu of a policy control backdrop. However, evaluating minimization techniques against the utility goals in Table 2 show the weakness of this one-dimensional technical approach. This weakness is unsurprising, since data minimization techniques intentionally obfuscate information often essential to most Internet research. These utility gaps can be modulated ("dialed down") with the policy components of PS2. In short, a purely technical approach breaks down along the utility dimension, and the pure policy approach may leave too high privacy risk exposure, justifying a hybrid framework that covers both privacy risks and utility goals. We note that evaluation of a framework must also consider practical issues such as education costs, whether new privacy risk(s) are introduced, whether control(s) are forward-looking or also address legacy privacy risks, and free rider problems created by DPs who choose not to share.

PS2 Privacy Risk	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. The X's indicate that the particular PS2 component in the row fails to mitigate against the privacy risk enumerated in the corresponding column. (*Minimization* refers to the techniques in Table 2.)

Minimiz.Tech.	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.

References

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

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

[3] Burkhart, M., Schatzmann, D., Trammel, B., Boschi, E., and Plattner, B. The role of network trace anonymization under attack. *ACM SIGCOMM Comp. Comm. Rev.* (2009).

[4] Burstein, A. Amending the ECPA to Enable a Culture of Cybersecurity Research. *Harvard Journal of Law & Technology 22*, 1 (2008), 167-222.

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

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

[7] Center, E. P. I. The U.S. First Circuit Court of Appeals upheld a New Hampshire law that bans the sale of prescriber-identifiable prescription drug data for marketing purposes. `http://epic.org/privacy/imshealth/11_18_08_order.pdf`.

[8] Center for Democracy and Technology. CDT's Guide to Online Privacy, 2009.

[9] Clabby, C. DNA Research Commons Scales Back. *American Scientist 97*, 3 (May 2009).

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

[11] Crovella, M., and Krishnamurthy, B. *Internet Measurement: Infrastructure, Traffic and Applications*. John Wiley and Sons, Inc., 2006.

[12] Narayanan, A., and Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on Security and Privacy* (2008).

[13] OECD. Guidelines on the protection of privacy and transborder flows of personal data, 1980.

[14]

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 (2008).

[15]

Swire, P. A Theory of Disclosure for Security and Competitive Reasons: Open Source, Proprietary Software, and Government Agencies. *Houston Law Review 42*, 5 (January 2006).

Dialing privacy and utility: a proposed data-sharing framework to advance Internet research

Erin E. Kenneally and Kimberly Claffy
Cooperative Association for Internet Data Analysis (CAIDA)
University of California, San Diego *
erin,kc@caida.org

1. INTRODUCTION

We re-visit the common assumption that privacy risks of sharing 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. This window of opportunity lies in public re-examination of the assumption that the privacy risks of sharing network measurement data outweigh the benefits, and for stakeholders to self-regulate in the interests of building social capital and informing legal and judicial regimes. 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).

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. The threat model from not data sharing is necessarily vague, as damages resulting from knowledge management deficiencies are beset with causation and correlation challenges. More fundamentally, 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. Admittedly, 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 the information economy it supports.

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 can effectively manage privacy risks that have heretofore impeded more than ad hoc or nod-&-a-wink data exchanges. Our model integrates privacy-enhancing technologies with a policy framework that applies proven and standard privacy principles and obligations of data seekers and data providers, in coordination with techniques that implement and enforce those obligations. We evaluate this framework along two primary criteria: (1) how they well the policies and techniques address privacy risks; and, (2) how well policies and techniques achieve utility objectives. We also include a case study showing how we apply the principles and techniques of the framework to share network operational data for use in cybersecurity R&D.

2. CHALLENGES AND MOTIVATIONS

Historically, Internet data of interest to network researchers has included IP topology data, traffic traces including traffic to unused address space, full packet captures of DDOS, worm, or botnet communications, exported flow records, and exterior and interior routing table data [8]. Our collective use of and dependence on the Internet continually grow, and accordingly so does the range of disciplines which must study aspects as scientifically as possible. An expanding range of Internet data of potential interest and utility to an expanding domain of researchers must bring with it deeper consideration of privacy in the collaboration and data sharing models, especially between industry and academia.

The strategic challenge is similar to other domains: how to balance utility goals with privacy risks for data seekers (DS) and data providers (DP). Internet researchers and systems security personnel are generally DS – entities seeking to share, responsibly disclose, acquire or otherwise exchange real world data. 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

*This work is sponsored by the U.S. Department of Homeland Security (DHS) Science and Technology (S&T) Directorate NBCHCC040159.

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.

Implicit incentives to share measurement data exist, but their 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, ad hoc and/or opaque exchange protocols exist, but measuring their effectiveness and benefit is challenging. Consequently it is difficult to justify resources for the research and collaboration costs that incentivize a sharing regime. On the other hand, the high cost of independently acquiring datasets is a motivation for re-use where possible. Transactional and opportunity costs include administrative and legal permissions, hardware and software support for instrumentation, and human capital to monitor instrumentation, and manage and curate data [1].

Privacy is difficult to quantify, as is the utility of measurement-based research. Both variables are dynamic and lack normative understanding among both domain professionals and the general citizenry. As fields of study, privacy and network science are both hindered by the absence of: common vocabulary, open and configurable reference models, uniform means of analysis, common sets of use cases, and unsurprisingly, any standard cost (liability) accounting or ROI formulas. A circular conundrum 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. While there are no procedures or regulatory framework to foster widespread exchange, there is also no framework that prohibits it either. Although there is interest in efficient and widespread 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

Under the U.S. and European Union (EU) regulatory regimes, the concept of personally identifiable information (PII) is central to privacy law and data stewardship. It is commonly is defined as information which can be used alone or in combination with other information to distinguish or trace an individual's identity[10]. Unlike the EU model which allots overarching protection for PII, the U.S. protects PII across a patchwork of caselaw, state and federal industry-specific laws covering health data, financial data, education data, employment data, insurance records, government-issued records, credit information, and cable and telephone records.

At its core, there is ambiguity over fundamental concepts upon which privacy risk assessment turns. First, 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 that the concept of PII in Internet data is not clear – its definition is context-dependent, both in terms of technology and topology. Further, the ability to link network data to individuals – as well as the cost of doing so – changes over time as technologies and protocols evolve. Yet, PII is fundamental to interpreting and applying many laws.

For example, binary and blanket characterizations of Internet protocol addresses (IPA) or uniform resource locators (URLs) as PII (or not) are necessarily inaccurate because neither alone can capture the range of privacy risks. In practice there is little functional differentiation between these traffic components and other, traditionally protected PII, yet the related legal treatment of IPAs and URLs is far less consistent. A more accurate risk assessment depends on context, i.e, metadata about who collected it and how they use, disclose, and dispose of the traffic data.

The risk management challenge lies in the linguistic incongruity between the legal and technical discourse about traffic data – its definitions, semantic classifications and interpretations. Officers of the court often associate IPAs with a greater privacy risk than URLs based on our ability to link IPAs to an individual via service provider account records. This distinction is artificial (albeit not totally 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.

Furthermore, this legal-technical gap exposes privacy risks with network operational data insofar as many laws do not explicitly allow for research use of network data [4], and there is no bright line caselaw applying their respective exceptions to the context of sharing Internet data for research.

2.2 Flawed Technology Models

Most data-sharing efforts by the networking research community focus on improving privacy enhancing technologies (PET) to solve the privacy problem, with anonymization commanding the bulk of the attention. A typical research approach is to enumerate the possibly privacy-sensitive information present in network traffic traces, and then implement a technical, typically cryptographic,

solution to replace this information completely or partially with synthetic identifiers, normally implemented by encrypting or otherwise removing all or part of identifiers.

Since privacy risk is influenced by evolving contexts associated with relationships between people, data, technology and institutions, solely technical solutions are inherently insufficient to balance the privacy/utility tradeoff. Technical researchers may rightly ask why we would predicate sharing architectures on ambiguous, unquantifiable and fallible human trust enforced by law and policy, if we can build trust through technology. The response is simple: while a purely technical approach may significantly ameliorate privacy risk, it largely fails to render empirically grounded answers to most questions being asked about the Internet today.

For example, while anonymization schemes can enhance the privacy of IPA in shared network traces, if it removes the ability to do any sort of geographic or topological analysis, the research utility of that data for studying DDoS modus operandi is dramatically reduced. A policy control framework enables the technical dials to allow for more privacy risk if a specific use justifies it. For example, traces protected by prefix-preserving anonymization may be subject to reidentification risk or content observation risk, but policy controls can help data providers minimize the chances that sensitive information is misused or wrongfully disclosed [3].

2.3 Reactive Top-Down Policy

Strategies to incentivize sharing by amending or enacting legislation merit consideration. However, regulation, especially in the technology arena, is largely reactive rather than making proactive, fundamental adjustments to predictable difficulties. Further, the length of the legislative policy cycle, confluence of variables involved in changing law, and unpredictable change agents are not amenable to immediate solutions that interested stakeholder DS and DPs can execute. A legislative solution means awaiting the infamous change agent: body counts or billion dollar losses that result from the lack of ground truth about the structure and function of networks that comprise our critical communications infrastructure.

3. SHARING RISKS AND BENEFITS

3.1 Who, What, When

Who is at risk when network data is shared?
Entities potentially at risk when network traffic is shared include: persons who are identified or identifiable in network traffic, researchers, and network providers (NP) such as ISPs, backbone providers, and private network owners. In addition to legal liabilities and ethical responsibilities, researchers and their institutions also risk withdrawal of data and/or funding as a result of privacy leakage. Society also bears costs associated with misinformation, mistrust, and internalizing behavioral norms that may result from privacy harms.

Which traffic data components are privacy-relevant?
We call a *first-order identifier* one which functionally distinguishes an individual: first and last name, social security number, government-issued and other account identifiers, physical and email addresses, certain biometric markers, and possibly the same information about immediate family. A *second-order identifier* could be an IPA, machine access code (MAC) address, host name, birthdate, phone number, zip code, gender, and financial, health, or geographic information. These indirect identifiers can also include aggregated or behavioral profile information such as IP header information, which can reveal the applications used, how often, and from which machines. Indirect identifiers also include URL click streams, which can reveal information about the content of communications, including search terms.

Under what conditions do these data types pose risk?.
Network traffic data can present a privacy risk when information in packets and flow records can directly expose non-public information about persons – such as health, sexual orientation, political affiliation, religious affiliation, criminal activity, associations, behavioral activities, physical or virtual location; or, organizations – such as intellectual property, trade secrets or other proprietary information. Indirect risk exposure can occur when data is correlated (linked) with other public or private data, such as the case with IPAs of worm-infected and thus vulnerable hosts. Network data can also yield mistaken attributions and inferences about behavior.

The privacy risk across time may also vary – the threat may be immediately manifest upon disclosure of data, or it may may be a latent risk which is held in abeyance until some future condition arises. Lack of transparency between the DP and DS regarding the shared data's nature, scope, and lineage is invariably a condition that enhances risk.

3.2 Privacy Risks of Internet Research – Laws and Courts of Public Opinion

It is impractical to enumerate all laws that may affect privacy risk, but such inventorying is not prerequisite to capture the foreseeable risks of network data sharing. It is sufficient to note that legal liability or ethical obligations underlie each privacy risk. In the U.S., privacy-related legal liabilities can derive from the federal Constitution (most notably the Fourth Amendment), federal law and regulation, contract law, tort law (e.g., invasion of privacy), state law equivalents, and organizations' privacy policies. Beyond legal risks, violations of ethical obligations can create normative harms that

implicate reputation and financial damages. Dismissing ethical obligations as discretionary and unenforceable overlooks how ethical violations are treated by public opinion, and also ignores the fact that many laws are informed by ethical norms [7].

Public disclosure is the act of making data readily available to the general public via publication or posting on the web. The privacy risks of sharing data containing PII which is subsequently displayed on the web are obvious and incontrovertible. More common and challenging are publicly available network traces and activity logs which reveal identifying information about infected hosts. Such disclosure raises the risk that unpatched or vulnerable hosts will be further exploited, thus creating security and reputation risks for individuals and organizations.

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. AOL provided a quintessential example in 2006 when they released an anonymized data set of search queries that were linked back to users conducting the searches using public metadata. These persons were then exposed in the NYT. [2]

Compelled disclosure to third parties risk arises with the obligations attendant to controlling data, such as having to respond to subpoenas requesting data disclosure in lawsuits. The RIAA campaign to massively subpoena ISPs and universities in an attempt to identify copyright infringers is a notorious example. To avoid such risk, many entities (including research organizations) have chosen not to retain data, thereby also losing operational and research value.

Government disclosure involves the release of data to government entities. An infamous example is the disclosure of call data records by major telecommunications carriers to the National Security Agency around 2007 [6]. Release to the government introduces another level of risk involving civil rights and liberties, such as imprisonment and restrictions on speech and associations.

Misuse of user or network profiles arises with network traffic that contains information about proprietary or security-sensitive network architectures or business operations. Advancing traffic and topology analysis, data mining and classification techniques can derive sensitive information from seemingly benign traffic data, and thereby reveal user behaviors, associations, preferences or interests, which attackers, advertisers, or content owners can then exploit. Network operators themselves may use such information for network management, illustrated by Comcast's recent throttling of BitTorrent traffic.

Inference misuse risk involves synthesizing first-order or second-order identifiers to draw inaccurate inferences about a person's behavior or identity that leads to damage or harm.

Re-identification /De-anonymizing misuse risk. Re-identification/de-anonymization, involves reversing data anonymization or masks to link an obfuscated identifier with its associated person. Shared anonymized data poses a misuse risk because it is variably vulnerable to re-identification attacks using public or private information whose availability is beyond the knowledge or control of the original or intermediate data provider [12]. Anonymized data may not immediately expose PII, but any time a piece of de-identified data has been linked to first order identifying information, other anonymous aspects of the obfuscated data are easier to de-anonymize. Aggregation or statistical techniques for anonymization are not immune to re-identification risk.

Examples of reidentification risk are the 2007 Netflix prize incident [9], and a similarly embarrassing episode of re-identification within the Internet research community [1].

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 turn to anonymization to avoid triggering privacy laws that exempt aggregate or anonymized data.

Like the arms race between exploits and defenses in the systems security arena, de-anonymization techniques will likely become commoditized to support investigative reporting, law enforcement, business intelligence, research, legal dispute resolution, and the presumed criminal threatscape.

3.3 Utility of Internet Measurement

The benefits of network research derive from the value of empirical network science [5], which includes a better understanding of the structure and functions of networks that comprise critical Internet infrastructure. Network researchers and funding agencies struggle to establish a science agenda, partly due to their lack of visibility into the infrastructure, but also because the field is younger and less well-defined than traditional scientific disciplines.

The following are criteria against which to measure, evaluate and communicate the benefits of sharing network data for research:

- The objective for sharing the data promotes social welfare or generalizable knowledge.
- The data is not already being shared, or if it is, there remains a qualitative need for sharing be-

tween other DS and DPs
- The research could not be conducted without the data.
- The scientific methodology using the data is transparent, objective, and repeatable relative to any privacy controls that are implemented.
- Research results can be acted upon meaningfully.
- Research results can be integrated with business processes, such as situational awareness of critical infrastructure or operational security.

Research in network measurement that could satisfy the above criteria include:

- information and network security questions regarding system threats, including characterizing baseline and anomalous workloads, modeling malware, developing effective strategies to deal with threats.
- macroscopic analysis of Internet topology; understanding the how the evolution of the network is affecting the efficiency and capabilities of the underlying routing, transport, and naming protocols.
- understanding the effect of the prevalence and growth of new applications on Internet workload, topology, and infrastructure economics.
- validation of traffic, congestion control, and performance assumptions, models, and analyses, both for current and proposed new technologies.
- development and evaluation of new tools and algorithms, including measurement and sampling techniques.

4. PS2 FRAMEWORK: ELEMENTS, EXECUTION, AND EVALUATION

We describe the Privacy-Sensitive Sharing Framework and then evaluate the model's ability to address the privacy risks outlined in 3.2 and the utility criteria in 3.3. Recognizing that privacy risk management is a collective action problem, our PS2 framework contains this risk by conveying the collection, use, disclosure and disposition controls over to the DS coincident with the shared data. This framework contemplates that the privacy risks associated with shared data are contagious – if the data is transferred, some degree of responsibility for containing the risk lies with both provider and seeker of data.

4.1 Elements of PS2

The PS2 is a structured framework for describing privacy risks and controls to support and implement functional privacy requirements. It serves three purposes: an analytical tool for assessing the risk posture of the proposed data sharing; a basis for establishing privacy management (technical and policy) controls; and a template for developing operational solutions to balancing privacy and utility in data sharing.

While not anchored on specific regulation, the components of our framework are rooted in principles and practices that underlie privacy laws and policies on both the national and global levels. 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 [11]. The PS2 framework – a hybrid of policy and technical controls – applies these principles to the context of Internet research, allowing navigation of data disclosure and misuse risks, and serving as a touchstone for legally and ethically defensible data sharing.

- Authorization – Internal authorization to share requires explicit agreement between the DP and DS. This may require direct consent from individuals identifiable in network traffic or via proxy consent with the DP.[1]
- Oversight – The DP and DS should engage external oversight of the proposed sharing, such as from an Institutional Review Board (IRB).
- Transparency – The DP and DS should be open and in agreement over the collection, use, disclosure, objectives and obligations and associated with shared data. For example, data-sharing terms might require that the algorithms be public but that the data and/or conclusions remain protected, or vice versa [13].
- Compliance with applicable law(s) – Collection, use and disclosure of data should comport to a reasonable if not case-law precedented interpretation of laws that speak directly and clearly to sharing risks about proscribed behaviors or mandated obligations.
- Purpose adherence – The data should be used consistent with the documented goal for why it is being shared.
- Access limitations – The shared data should be restricted from those who do not have a need and right to access the shared data.
- Use specification and limitation – Unless otherwise agreed, the DP should prohibit merging or linking data that would create or enhance privacy risk.
- Collection and Disclosure Minimization – The DP should collect and disclose only the data that is necessary to achieve the research goals, and eliminate extraneous data that carries a privacy risk. Prominent privacy-sensitive techniques include:
  A. Deleting/filtering sensitive data.
  B. Deleting/filtering part(s) of the sensitive data.
  C. Anonymizing/hashing/de-identifying all or parts of the sensitive data.
  D. Aggregating or sampling.

[1]Consent requirements for Internet traffic monitoring are unresolved, but will no doubt be a part of forthcoming legal, policy and community decisions.

E. Mediation analysis /human proxy – this is a sandbox approach that involves "sending the code to the data" rather than releasing sensitive data for analyses.
F. Aging the data – traffic data that is de-sensitized by virtue of being non-current, i.e., no longer contains a direct or indirect identifier that poses a risk of harm.
G. Size/quantity limitation – this entails minimizing the quantity of traces shared.
H. Multiple layers of anonymization.

- 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 – Procedures to address harms from inappropriate data use or disclosure, including a feedback mechanism to support correction of datasets and/or erroneous conclusions.
- Data and analysis quality assurances – Awareness by the DS and DP of inference confidence levels associated with the data.
- Security – Controls should reasonably ensure that sensitive PII is protected from unauthorized collection, use, disclosure, and destruction.
- Training – Those who are authorized to engage the data should be educated and made aware of the privacy principles and controls associated with the data.
- Impact assessment – Sharing dynamics should consider potential collateral effects on stakeholders affected by the data, and seeks methods that do no further harm.
- Transfer to third parties - This should be prohibited unless equivalent data control obligations are transferred, relative to the disclosure risks associated with that data.

4.2 Execution of PS2

The technical controls of the PS2 are self-contained, although they need to be identified and enforced. The policy controls require an execution vehicle, such as a bi/multilateral Memoranda of Understanding (MOU) or Memoranda of Agreement (MOA), a model contract, or a binding organizational policy. For lower risk sharing situations, a unidirectional Acceptable Use Policy AUP may be cost-preferential to negotiated bilateral agreements. Mutual and explicit consent to engage policy and technical controls provides an enforceable standard and certainty that can serve as a safe harbor for liability under many data privacy laws.

4.3 Evaluation of PS2

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

Table 1: PS2 policy and technical components evaluated against privacy risks. The ✓'s indicate that the particular PS2 component in the row addresses the privacy risk enumerated in the corresponding column. (*Minimization* refers to the techniques in Table 2.)

The PS2 framework offers a template for assessing and developing operational solutions for balancing privacy risks and utility rewards when sharing data for research. It can help an oversight committee determine whether possible risks are justified, by specifically 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 the prospective DP, the assessment will assist the determination whether or not to participate.

Table 1 illustrates whether the privacy risks are mitigated by the primary components of PS2. The ✓'s indicate that the particular PS2 policy component in the row mitigates against the privacy risk enumerated in the corresponding column. The table illustrates that the policy control component of PS2 leaves gaps in addressing the full range of privacy risks. Further, it suggests that the technical control component (minimization techniques) (Section 4.1) can, however, address all privacy risks. The implication is that a purely technical sharing framework is sufficient to address privacy risks, and therefore a policy control backdrop is superfluous. However, evaluating the technical minimization controls against the utility goals in Table 2 illustrates

Minimiz.Tech./Utility Need	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?
Not Sharing	X	X	X	X	X	X
Filter All	X	X	X	X	X	X
Filter Part	X	X		X	X	
Anonymize	X	X	X	X	X	
Aggregate	X	X	X	X	X	
Mediate (SC2D)	X					
Age Data	X	X	X	X	X	
Limit Quantity	X	X	X	X	X	
Layer Anonymization	X	X	X		X	

Table 2: PS2 technical component (minimization controls) evaluated against utility needs.The X's indicate where the minimization technique impedes the research utility goal.

the weakness of a one-dimensional technical approach. This weakness is unsurprising, since data minimization techniques intentionally obfuscate information often essential to most Internet research. These utility gaps can be modulated ("dialed down") by engaging the policy components of PS2. In short, a purely technical approach breaks down along the utility dimension, and the pure policy approach may leave too high privacy risk exposure, justifying a hybrid framework that covers both privacy risks and utility goals. We note that evaluation of the framework should also consider practical issues such as education costs, whether new privacy risk(s) are introduced, whether control(s) are forward-looking or also address legacy privacy risks, and possible free rider problems created by DPs who choose not to share.

5. PS2 CASE STUDY: NETWORK TELESCOPE

To promote cooperative analysis of Internet traffic and performance and advance the state of cybersecurity research, we have implemented the policy-supported and risk-sensitive PS2 data sharing framework. We recently applied this framework to a then-new mode of data sharing: real-time sharing of Internet traffic data observed at the network telescope. A network telescope is a segment of routed IP address space on which little or no legitimate traffic exists. Each such chunk of address space provides a unique and continuous view of anomalous, unsolicited Internet traffic to no legitimate destination.

Observing traffic from a network telescope allows visibility into a wide range of security-related events, including misconfiguration (e.g. a human being mis-typing an IP address), malicious scanning of address space by hackers looking for vulnerable targets, backscatter from random source denial-of-service attacks, and the automated spread of malicious software called Internet worms. The primary obstacles to sharing telescope data are privacy and security concerns. Because viruses and worms may involve the installation of backdoors that provide unfettered access to infected computers, telescope data may advertise these vulnerable machines.

CAIDA previously addressed the privacy risk of releasing victim host IPAs and unexpected but occasional payload content with strict filtering and anonymization disclosure controls. This implementation of privacy risk controls came at research utility costs, which two events in 2009 motivated us to re-examine: the Conficker worm outbreak, and a new storage cost allocation structure in our organization. In 2009 we transitioned from a model of static trace sharing and indefinite storage of data on CAIDA servers, to a model of real-time data sharing with vetted researchers, storing only a 30-day window of history. This new model aims to allow researchers access to a telescope observatory during a worm outbreak, where raw traces containing target addresses and payload that could enable autopsy of the structure and function of cybersecurity threats.

Consistent with the PS2 framework, we use Acceptable Uses Policies (AUP) and disclosure control techniques to guide this shift in our data-sharing approach. We implement transparency by clearly describing the dataset and its use obligations on our public website. We obtain explicit consent to abide by the stated responsibilities by requiring each researcher (DS) to complete and execute a data request form which includes acknowledging data use terms prior to receiving access. External oversight is addressed by our university's Institutional Review Board, which certifies that the datasets are collected and made available in accordance with the principles of respect for persons, beneficence and justice as relevant to human subjects. We institute purpose specification by obtaining explicit webform acknowledgment from the DS that s/he will use the dataset solely for the stated research purposes. We enforce access to the dataset(s) and authorization to use it by application review, approval and communication of acquisition instructions by CAIDA administrators. This review includes restricting access to DS from export-restricted countries. DS also consent to use appropriate and reasonable care in safeguarding access to and preventing unauthorized use of the data. We obtained legal advice to ensure sharing methods comply with laws and policies related to data privacy, confidentiality and pro-

tection.

Another element of the PS2 is the impact assessment. CAIDA researchers and administrators considered possible harm to individuals or organizations, as well as the likelihood of achieving the growing needs of security research by only releasing and storing completely sanitized versions of static, periodic datasets. To the extent possible within the real time strategy, privacy sensitivities were addressed with loosened disclosure controls (anonymization of any identifying payload). Our AUP backstopped re-identification risk by requiring that researchers agree to make no attempts to reverse engineer, decrypt, or otherwise identify the original IP addresses collected in the trace.

As discussed in 4.3, our original disclosure control strategy largely ameliorated the privacy risks of disclosing victim IPAs and payload, but to the detriment of security research utility needs. The speed, scope, and strength of automated malicious software demand effective real-time sources of data that matches the dynamics of the threat. Studying a worm in situ requires real time traffic access, including raw victim host IPAs, and payload data. None of these needs were supported by our original disclosure control strategy.

The PS2 hybrid framework allowed CAIDA to realize utility goals in a risk-sensitive manner, by dialing down the technical disclosure controls, and relying on policy components to close resulting privacy gaps. Specifically, we collected and shared telescope data as raw (unanonymized) traces, with payload (content). Rather than mitigate the risk of releasing victim IPA by anonymization and wholesale deletion of security-relevant data, CAIDA revisited the privacy impact assessment, loosened the technical disclosure controls, and tightened its use and disclosure obligations in the AUP. CAIDA used the considerations enumerated in Table 2 and Section 3.3 to enable transparent and reproducible scientific research of a critical infrastructure security event *while still in progress*, a contribution to the security field not possible with previously available data sets.

The privacy risks associated with this dataset were such that CAIDA could effectively manage them by relaxing the technical anonymization barriers to research utility, and constricting DS use and disclosure via the policy component. For example, the DS are prohibited from attempting to connect to, probe, or in any other way interacting or intervening with a machine or machine administrator identified in the dataset, without permission from CAIDA. For any publication or other disclosure of non-anonymized data, the DS is obligated to anonymize or aggregate IP addresses, network names, and domain names unless obtaining written authorization from CAIDA to do otherwise. We (do our best to) enforce compliance with these restrictions with an audit policy that requires the DS to report a summary of the research and any findings, publications, or URLs using the data to CAIDA at the conclusion of the research, or semi-annually.

6. CONCLUSIONS

The Privacy Sensitive Sharing (PS2) framework considers practical challenges confronting security professionals, network analysts, systems administrators, researchers, and legal advisors. It embodies the proposition that privacy problems are exacerbated by a shortage of transparency surrounding the who, what, when, where, how and why of sharing privacy-sensitive information. The PS2 enables transparency as a touchstone of data-sharing.

The PS2 offers a consistent, transparent and replicable evaluation methodology for risk-benefit determinations rather than relying on subjective, opaque and inconsistent evaluations that turn on *trust-me* decision metrics. The PS2 is a hybrid approach: a policy framework that applies proven and standard privacy principles between the data seekers and data providers, coordinated with technologies that implement and provably enforce those obligations. We evaluated this framework along two main criteria: (1) how well the policies and techniques address privacy risks; and, (2) how well policies and techniques achieve utility objectives.

We hope this framework helps network measurement advocates use this window of opportunity to experiment with models that effectively manage privacy risks which have heretofore impeded more than ad hoc or nod-&-a-wink data exchanges. Because the principles underlying data stewardship are domain-agnostic, the PS2 principles can also help prevent a proliferation of infamous poster cases [2, 9, 1] across disciplines.

By taking proactive and ethically defensible steps to transparently engage sharing models like PS2, we can more practically influence policy and law at these crossroads. The alternative is to wait for a legislative reaction to catastrophe, at which time a window of opportunity will have closed. Information security controls were initially considered a liability (from a cost perspective) until regulations rendered lack of security a compliance liability. We anticipate circumstances to reveal that rather than data-sharing being a risk, *not* sharing data is a liability. We offer the PS2 as a tool to help move the community mindset in that direction as productively and safely as possible.

7. REFERENCES

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

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

[3] Burkhart, M., Schatzmann, D., Trammel, B., Boschi, E., and Plattner, B. The role of network trace anonymization under attack. *ACM SIGCOMM Comp. Comm. Rev.* (2009).

[4] Burstein, A. Amending the ECPA to Enable a Culture of Cybersecurity Research. *Harvard Journal of Law & Technology 22*, 1 (2008), 167–222.

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

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

[7] Center for Democracy and Technology. CDT's Guide to Online Privacy, 2009.

[8] Crovella, M., and Krishnamurthy, B. *Internet Measurement: Infrastructure, Traffic and Applications*. John Wiley and Sons, Inc., 2006.

[9] Narayanan, A., and Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on Security and Privacy* (2008).

[10] National Institute of Standards and Technology. Guide to Protecting the Confidentiality of Personally Identifiable Information, January 2009.

[11] OECD. Guidelines on the protection of privacy and transborder flows of personal data, 1980.

[12] 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 (2008).

[13] Swire, P. A Theory of Disclosure for Security and Competitive Reasons: Open Source, Proprietary Software, and Government Agencies. *Houston Law Review 42*, 5 (January 2006).

The Cooperative Association for Internet Data Analysis

www.caida.org > data : : overview

CAIDA Data - Overview of Datasets, Monitors, and Reports

CAIDA collects several different types of data at geographically and topologically diverse locations, and makes this data available to the research community to the extent possible while preserving the privacy of individuals and organizations who donate data or network access.

This page provides a quick-access overview of available datasets (publicly available or otherwise restricted), with links to the dataset descriptions and access request forms when applicable. To read more about CAIDA's efforts in data curation and promoting data sharing, see the main CAIDA Data page.

Recent Datasets:

- DDoS Attack 2007 Dataset (2010-02-23)
- Three Days of Conficker Dataset (2009-09-02)
- Two Days in November 2008 Dataset (2009-07-01)
- IPv6 Topology Dataset (2009-03-20)

The colored header in the overview table below indicates the field(s) the table is being sorted by. Click the header to change or toggle sorting order, and Shift+click headers to sort by multiple fields. Rows are colored by the collection status of the indicated dataset as follows:

Ongoing	The data collection for this dataset is still active and has continuing, regularly scheduled collections.
One-time snapshot	The dataset comes from a single collection event that only occured once. Future events will have a different dataset name.
Complete	A formerly ongoing data collection that is finished, and will not be resumed.

Name	Status	Availability	Category	Source	Anonymization	Release Date
AS Links: IPv4 Routed /24 AS Links	Ongoing	Public (download)	Topology	Ark + RouteViews	none	2007-09
AS Rank	Ongoing	Public (web query)	Topology	Ark + RouteViews	none	2004-01
AS Relationships	Ongoing	Public (download)	Topology	RouteViews	none	2004-01
DNS root/gTLD RTT Data	Ongoing	Request access (download)	Performance	NeTraMet	none	2006-08
Equinix Chicago A: Passive Realtime Monitor	Ongoing	Public (interactive graph)	Traffic Summary Statistics	OC192 Monitors	Crypto-PAn	2008-03
Equinix Chicago B: Passive Realtime	Ongoing	Public (interactive graph)	Traffic Summary Statistics	OC192 Monitors	Crypto-PAn	2008-03

Name	Status	Availability	Category	Source	Anonymization	Release Date
Monitor						
Equinix San Jose A: Passive Realtime Monitor	Ongoing	Public (interactive graph)	Traffic Summary Statistics	OC192 Monitors	Crypto-PAn	2008-07
Equinix San Jose B: Passive Realtime Monitor	Ongoing	Public (interactive graph)	Traffic Summary Statistics	OC192 Monitors	Crypto-PAn	2008-07
Live Telesope	Ongoing	Restricted (Invitation only)	Traffic	Telescope	none	2009-03
Network Telescope: Passive Realtime Monitor	Ongoing	Public (interactive graph)	Traffic Summary Statistics	Telescope	none	2001-02
RouteViews Prefix to AS mappings	Ongoing	Public (download)	Topology	RouteViews	none	2005-05
SDNAP: Passive Realtime Monitor	Ongoing	Public (interactive graph)	Traffic Summary Statistics	SDNAP Monitor	none	1998-02
Topology: IPv4 Routed /24	Ongoing	Request access (download)	Topology	Ark	none	2008-09
Topology: IPv4 Routed /24 DNS Names	Ongoing	Request access (download)	Topology	DNS	none	2008-03
Topology: IPv6 Topology	Ongoing	Request access (download)	Topology	Ark	none	2008-12
DDoS Attack 2007	One-time snapshot	Request access (download)	Security	CoralReef	Crypto-PAn	2010-02
Telescope Data (Three Days of Conficker)	One-time snapshot	Request access (download)	Traffic	Telescope	destination network masked	2009-09
Telescope Data (Two Days in	One-time snapshot	Request access (download)	Traffic	Telescope	destination network masked	2009-06

Name	Status	Availability	Category	Source	Anonymization	Release Date
November 2008)						
Traceroute Probe Method 2008-08	One-time snapshot	Public (download)	Topology	Ark	none	2008-08
Witty Worm (public)	One-time snapshot	Public (download)	Worm Summary	Telescope	none	2004-03
Witty Worm (restricted)	One-time snapshot	Request access (download)	Worm Summary	Telescope	none	2004-03
Topology: Macroscopic Internet Topology Data Kit (ITDK)	Complete 2010-01	Request access (download)	Topology	Ark + RouteViews	none	2010-03
Topology: PAM 2010 "Improving AS Annotations"	Complete 2009-10	Request access (download)	Topology	Ark + RouteViews	none	2010-03
AS Links: Skitter (= AS Adjacencies)	Complete 2008-02	Public (download)	Topology	Skitter + routeviews	none	2006-05
Topology: Raw Skitter Topology	Complete 2008-02	Request access (download)	Topology	Skitter	none	2004-07
Code-Red Worms	Complete 2001-08	Public (download)	Security	Telescope	none	2001-07
Anonymized OC192 Traces 2008	Complete	Request access (download)	Traffic	OC192 Monitors	Crypto-PAn	2008-06
Anonymized OC192 Traces 2009	Complete	Request access (download)	Traffic	OC192 Monitors	Crypto-PAn	2009-03
Anonymized OC48 Traces (2002-2003)	Complete	Request access (download)	Traffic	OC48 Link	Crypto-PAn	2006-05
AS Taxonomy	Complete	Public (download)	Topology	RouteViews	none	2004-01
Backscatter 2004-2005	Complete	Request access (download)	Security	Telescope	destination network masked	2005-11

Name	Status	Availability	Category	Source	Anonymization	Release Date
Backscatter 2006	Complete	Request access (download)	Security	Telescope	destination network masked	2006-11
Backscatter 2007	Complete	Request access (download)	Security	Telescope	destination network masked	2007-11
Backscatter 2008	Complete	Request access (download)	Security	Telescope	destination network masked	2008-06
Backscatter TOCS	Complete	Request access (download)	Traffic	Telescope	destination network masked	2005-06
Router Adjacencies = Router Graph Links	Complete	Public (download)	Topology	Skitter	none	2003-04

Column Name	Column Description
Name	Dataset name
Status	Current status of this dataset, indicating whether the dataset collection is ongoing or not. "One-time snapshot" implies complete. A link to resulting papers or analysis would be linked from the status, if available.. • Ongoing - The data collection for this dataset is still active and has continuing, regularly scheduled collections. • One-time snapshot - The dataset comes from a single collection event that only occured once. Future events will have a different dataset name. • Complete - A formerly ongoing data collection that is finished, and will not be resumed.
Availability	Public datasets are downloadable per the Acceptable Use Policy for the dataset. Restricted datasets require prior permission as well as adherence to the Acceptable Use Policy for that dataset. If a dataset requires a formal request for usage, the form will be linked.
Category	• Security • Topology • Traffic • Traffic Summary Statistics • Worm Summary
Source	The source (e.g., network monitors, measurement infrastructure, etc) used to collect the dataset.
Anonymization	If the dataset is anonymized, the method of anonymization will be indicated. • Crypto-PAn - The Crypto-PAn tool was used to anonymize the IP addresses in the dataset. • destination network masked - The destination network was masked in the dataset.
Release Date	The date when the dataset or report was made available.

Additional Information

To keep up to date on CAIDA datasets you can subscribe to data-announce@caida.org. For other questions about CAIDA data, please contact data-info@caida.org. For more information about using CAIDA data, please see the CAIDA Data Usage FAQ.

Internet Measurement Data Catalog

Welcome to DatCat[SM]

The **DatCat** catalog indexes Internet measurement data.

DatCat lets you **find**, **annotate**, and **cite** data.

The goals of the system are:

- to facilitate searching for and sharing of data among researchers
- to enhance documentation of datasets via a public annotation system
- to advance network science by promoting reproducible research

For more information, see the general documentation.

News

2007-03-09 Added abilily to organize data by the Publications in which they are used, and to organize Collections of data into hierarchies (see the Object Types documentation).

2007-01-30 New features for contributors (see the contributing documentation).

2006-11-13 New feature: Detail pages include BibTeX citations

2006-10-20 New features: Advanced search interface is friendlier and more powerful; and any logged-in user can add a "note" annotation to any object.

2006-06-12 DatCat opened to public viewing

Catalog Statistics

object type	count	size

collection	106	—
publication	15	—
data	154 366	26.1 TB
package	127 904	13.9 TB

Browse or search the Catalog as a Guest

Log in

Create an Account

Sponsored by:

National Science Foundation

Software version 1.7.26
Page generated at 2010-03-20 21:10:10 UTC
Request processed in 0.0020 seconds

cooperative association for internet data analysis
based at the University of California's San Diego Supercomputer Center

DatCat℠

Internet Measurement Data Catalog

Collection: AOL 500k User Session Collection

Web queries to AOL search engine

Jump to: Description | Annotations | Citation | Record Details

Collection Contents

- data objects: 10 directly contained, 10 total

Collection Details

Summary	This collection consists of ~20M web queries collected from ~650k users over three months. The data is sorted by anonymous user ID and sequentially arranged. The goal of this collection is to provide real query log data that is based on real users. It could be used for personalization, query reformulation or other types of search research.
Motivation	This collection provides AOLs search query data files
Data Start Time	2006-03-01 00:01:03 UTC (+0000)
Data End Time	2006-05-31 23:59:59 UTC (+0000)
Data Duration	91 days 23:58:56 (7 948 736.0 s)
Creators	AOL 500k User Session Collection creator
Primary contact	*(none)*
Keywords	anonymized, AOL, search, search engine, search query, user modelling, web search
Used in publications	*(none)*
Member of collections	*(none)*

Description

This collection is distributed for NON-COMMERCIAL RESEARCH USE ONLY. Any application of this collection for commercial purposes is STRICTLY PROHIBITED.

Brief description:

This collection consists of ~20M web queries collected from ~650k users over three months. The data is sorted by anonymous user ID and sequentially arranged.

The goal of this collection is to provide real query log data that is based on real users. It could be used for personalization, query reformulation or other types of search research.

The data set includes {AnonID, Query, QueryTime, ItemRank, ClickURL}.

```
AnonID - an anonymous user ID number.
Query  - the query issued by the user, case shifted with
         most punctuation removed.
QueryTime - the time at which the query was submitted for search.
ItemRank  - if the user clicked on a search result, the rank of the
            item on which they clicked is listed.
ClickURL  - if the user clicked on a search result, the domain portion of
            the URL in the clicked result is listed.
```

Each line in the data represents one of two types of events:

```
1. A query that was NOT followed by the user clicking on a result item.
2. A click through on an item in the result list returned from a query.
```

In the first case (query only) there is data in only the first three columns/fields -- namely AnonID, Query, and QueryTime (see above). In the second case (click through), there is data in all five columns. For click through events, the query that preceded the click through is included. Note that if a user clicked on more than one result in the list returned from a single query, there will be TWO lines in the data to represent the two events. Also note that if the user requested the next "page" or results for some query, this appears as a subsequent identical query with a later time stamp.

CAVEAT EMPTOR -- SEXUALLY EXPLICIT DATA! Please be aware that these queries are not filtered to remove any content. Pornography is prevalent on the Web and unfiltered search engine logs contain queries by users who are looking for pornographic material. There are queries in this collection that use SEXUALLY EXPLICIT LANGUAGE. This collection of data is intended for use by mature adults who are not easily offended by the use of pornographic search terms. If you are offended by sexually explicit language you should not read through this data. Also be aware that in some states it may be illegal to expose a minor to this data. Please understand that the data represents REAL WORLD USERS, un-edited and randomly sampled, and that AOL is not the author of this data.

```
Basic Collection Statistics
Dates:
  01 March, 2006 - 31 May, 2006

Normalized queries:
  36,389,567 lines of data
  21,011,340 instances of new queries (w/ or w/o click-through)
   7,887,022 requests for "next page" of results
  19,442,629 user click-through events
  16,946,938 queries w/o user click-through
  10,154,742 unique (normalized) queries
     657,426 unique user ID's
```

Please reference the following publication when using this collection:

G. Pass, A. Chowdhury, C. Torgeson, "A Picture of Search"
The First International Conference on Scalable Information Systems, Hong Kong, June, 2006.

This data collection originally appeared at http://research.aol.com/pmwiki/pmwiki.php?n=Main.Research.

Annotations

Citation

Please use the following BibTeX citation to cite this collection. Some parts are **optional** or may need to be **edited**. To use the "`\url{...}`" command for nice URL formatting, you must call "`\usepackage{url}`" in the LaTeX preamble.

```
@MISC{/collection/1-003M-5=AOL+500k+User+Session+Collection,
  title = "{AOL 500k User Session Collection (collection)}",
  author = "{AOL 500k User Session Collection creator}",
  note = "\url{http://imdc.datcat.org/collection/1-003M-5=AOL+500k+User+Session+
    Collection} (accessed on 2010-03-20)",
  abstract = "Web queries to AOL search engine"
}
```

Record Details

Handle	imdc.datcat.org**/collection/1-003M-5**=AOL+500k+User+Session+Collection
Contributor	CAIDA Automated Data Contributor
Contributed	2006-08-09 19:55:13.179 UTC (+0000)
Last Modified	2006-08-09 19:55:13.179 UTC (+0000)

Software version 1.7.26
Page generated at 2010-03-20 21:18:05 UTC
Request processed in 0.058 seconds

cooperative association for internet data analysis
based at the University of California's San Diego Supercomputer Center