AN INTERNET DATA SHARING FRAMEWORK FOR BALANCING PRIVACY AND UTILITY

Erin Kenneally, M.F.S., J.D. kc Claffy, Ph.D.

Cooperative Association for Internet Data Analysis
University of California San Diego

Engaging Data | MIT | 12 October 2009

Talk Map

- Defining the Issue & Solution Space
- Value Proposition of PS2
- Challenges & Motivations
  - Uncertain Legal Regime
  - Incomplete Technology Solution Models
  - Privacy Risks
  - Under-valued Benefits of Network Measurement Research
- PS2 Framework
  - Policy Component
  - Technology Component
  - Implementation Vehicles
- Evaluating PS2
  - Privacy Risk Coverage
  - Utility Goals Coverage

The Issue Space Defining the Solution

- Issue Space
    - Current posture:
        - defensive, default-deny sharing network traffic data
    - (Misinformed) assumptions:
        - Privacy risks and legal restrictions >>> benefits of sharing
        - Unprecedented data availability = plethora of network infrastructure information
        - ISE directives post-911 → incent network data exchange
    - Muted legislative, judicial, policy drivers
        - Threat model from NOT sharing data = vague
        - No body count / $billion losses (at least no explicit, causal)
    - No widespread, standard procedures for exchange
        - Ad-hoc, nod & wink
    - Dynamic and normative-deficient understanding of privacy risk and research utility
        - No cost-accounting for privacy risk
        - No ROI for investment in empirical network measurement

    - Bright side of confusion = window of opportunity

Value Proposition of PS2

- Privacy-Sensitive Sharing (PS2) model solution

  = Privacy-enhancing technology + privacy-principled policies

- Risk – Benefit methodology
- Bridges risk – utility perception gap
- Enables transparency as touchstone of data sharing
- counter to subjective, opaque evaluations
- Engender trust, beyond “trust me”
- Considers practical challenges of stakeholders (network researchers, sys operators, security professionals, legal advisors, policymakers)
- Proactive, ‘self-regulation’
- Bottom-up regime
- Anchor point to demonstrate community norms, inform law & policy

Challenges & Motivations
(1) Uncertain Legal Regime

- No legal framework that explicitly prescribes, incentivizes, or forbids sharing of network data for security research
- Linguistic ambiguity between tech & legal discourse re: fundamental concepts driving risk
  - PII, Privacy, content, transaction data, URLs, IPAs, packet headers & body
  - Evolving tech increases capabilities and decreases costs of linking network data to individuals
  - Little functional difference between IPA, URL v. other protected PII, but law inconsistent
  - E.g., is IPA 'content' and URL 'addressing' data for ECPA and 4th A. purposes?
    - Johnson v. Microsoft (2008) - IPA does not identify persons
    - State v. Reid (2007) - REP in subscriber information attached to IPA
    - US v. Forrester (2007) - URLs may have REP because reveal communication content
    - HIPAA Privacy Rule – IPA is protected PII
    - States' data breach laws – IPA is not in definition of personal information
- Social normative expectations: my IPA, URLs + search terms are digital fingerprints?
  - Witness Tor, automated in-browser cookie and URL deletion

Challenges & Motivations
(2) Incomplete Technology Solution Models

- Point solutions fail to address context-dependent risks
  - Cases-in-point: de-anonymization attacks success
    - Prefix-preserving anonymization subject to re-identification
    - Poster cases
      - Netflix
      - Yahoo!
      - Traffic injection attacks
- Purely technical approaches necessarily impact research utility goals (analysis)
  - Data minimization techniques intentionally obfuscate essential data (network management, countering security threats, evaluating algorithms, apps, architectures)
  - E.g., Conficker

Challenges & Motivations
(3) Privacy Risks

- Derive from legal liabilities, ethical obligations, norms/court of public opinion
- 2 main categories
  - Disclosure risk
    - Public disclosure
    - Accidental/malicious disclosure
    - Compelled disclosure (e.g., RIAA subpoenas)
    - Government disclosure (e.g., NSA wiretapping, Telco releases)
  - Misuse risk
    - False inference (synthesizing 1st/2nd order identifiers to draw inferences about persons behavior, identity with damaging implications)
    - Network topology confidentiality
    - Re-identification/de-anonymization

      * increasing quantitatively & qualitatively
      - Cat & mouse game will drive commoditization of de-anon techniques
        - Pressure to protect (law, policy) AND motivation to uncover PII (profit, avoid legal liability triggers, attribution)
        - Law enforcement investigations, biz intel, legal dispute resolution, security incident response

Challenges & Motivations (4) Under-valued Benefits of Network Research

- Benefits:
  - Understanding structure, function of critical Internet infrastructure
  - (topology, workload, traffic routing, performance, threats & vulnerabilities)
- Network Data sharing utility criteria
  - Objective for sharing is positively related to social welfare
  - Need for empirical research
  - Research purpose not being conducted
  - Research could not be conducted without the shared data
  - No sufficiently similar data already being collected that could be shared
  - Research & peer reviewed methods using shared data are as transparent, objective, scientific and control for privacy risk
  - Results using shared data can be acted upon meaningfully
  - Results using shared data are capable of being integrated into operational or biz processes (security improvements, situational awareness)

PS2 Framework
Policy Components

- Core underpinnings:
  - privacy risks are 'contagious' (sharing= data AND responsibilities & obligations)
  - Components rooted in principles and practices of national & global laws, policies
    1. Authorization
    2. Transparency
    3. Compliance with applicable laws
    4. Purpose adherence
    5. Access limitations
    6. Use specifications and limitations
    7. Redress mechanisms
    8. Oversight
    9. Security
    10. Audit tools
    11. Data quality assurances
    12. Training
    13. Transfer to 3rd parties
    14. Ethical impact assessment
    15. Disclosure minimization

PS2 Framework
Technology Component

- Disclosure Minimization/Controls
  a) Deleting all sensitive data
  b) Deleting part(s) of sensitive data
  c) Anonymizing/de-identifying all or parts of sensitive data
  d) Aggregation or sampling techniques
  e) Mediation techniques (sending code-to-data)
  f) Aging the data
  g) Limiting quantity of data
  h) Layering anonymization

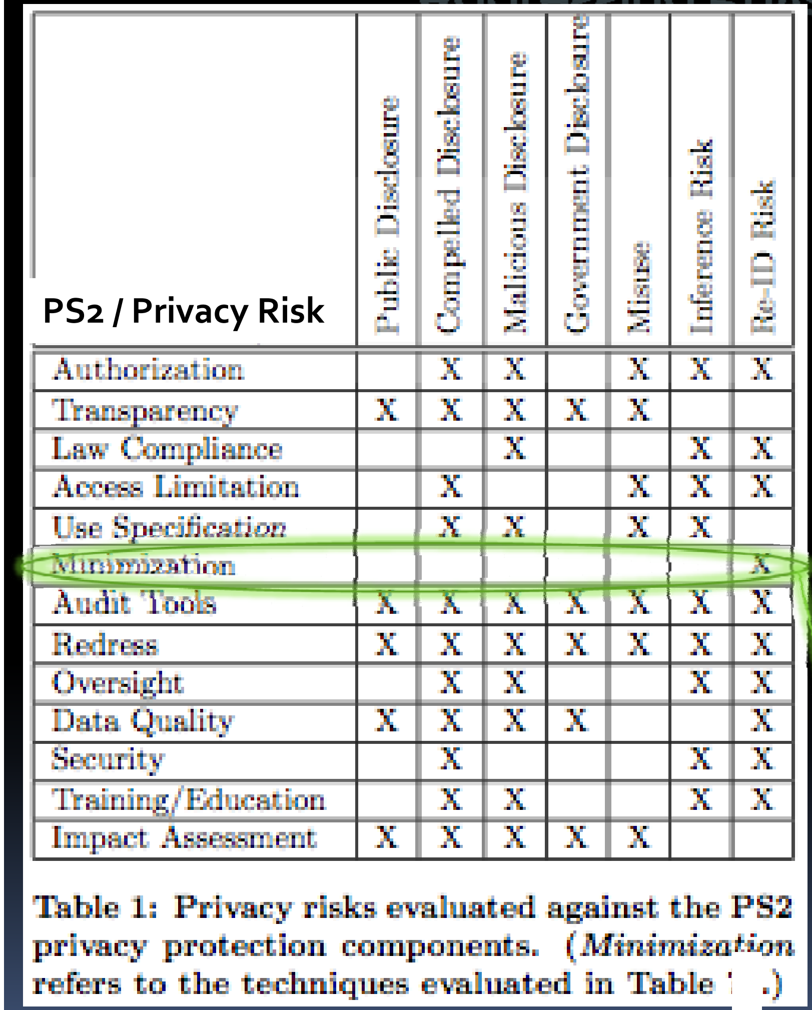
- Vehicles for Implementing PS2:
  - enforcement via MOU/MOA, model contracts, binding organizational policy, NDA

Evaluating PS2
Addressing Privacy Risk & Utility Goals

- Criteria:
  - 1. How well PS2 addresses privacy risks (table 1)
    - Policy control components, alone, leave coverage gaps
    - Technical controls, alone, seemingly control for privacy risks (implying policy control components superfluous)
  - 2. To what extent PS2 impedes utility goal (table 2)
    - Technical controls, alone, leave impedes utility

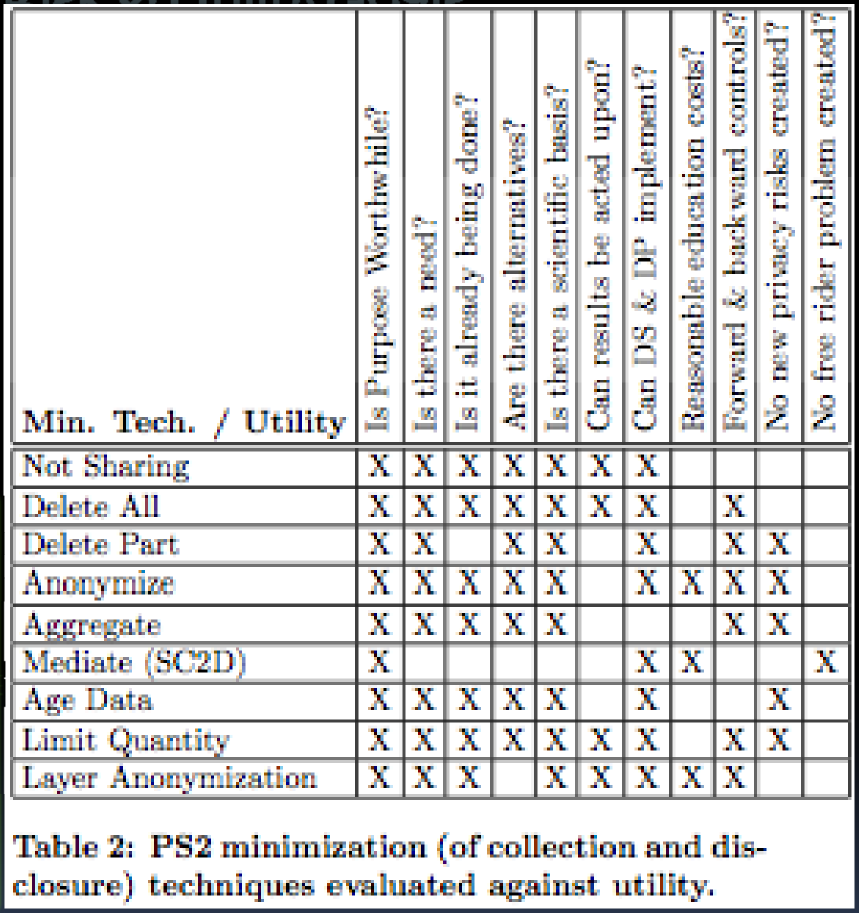
- Conclusion:
  - Singular tech solution breaks down along utility dimension
  - Singular policy solution leaves too high privacy risk exposure
  - Therefore, hybrid strategy allows tuning down technical controls to achieve utility objectives AND supplementing policy controls with preventative technical controls
  - Framework is both
    - Evaluation of hybrid model
    - Possible self-assessment tool for data sharing

Evaluating PS2
Addressing Privacy Risk & Utility Goals



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. (*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.

Bigger Picture:

Infosec controls evolved : financial liability ---> compliance necessity

PS2 value prop : regime where NOT sharing data ---> liability

go raibh maith agat

erin@caida.org