Approaches to enhancing privacy when conducting data science about people

Case Study using Record Linkage

Hye-Chung Kum, Associate Professor (kum@tamu.edu)
Population Informatics Lab (https://pinformatics.org/)
Department of Health Policy and Management, School of Public Health
Department of Computer Science and Engineering
Department of Industrial and Systems Engineering
The Center for Remote Health Technologies and Systems (CRHTS)
Texas A&M University
Data Science & Information Privacy
Barrier: Data Sharing
How do we conduct responsible research?

- Human interaction required for high quality data
  - Concerns about privacy
- A holistic approach
  - Data governance
  - Public engagement
  - Technical

Goal: To design an information system that can enforce the varied continuum from one end to the other such that one can balance privacy and usability as needed to turn data into decisions for a given task.
The start ...

- Write up a research plan on
  - What data you need
  - What do you want to do with them
  - Determine access levels for each data
- Submit to IRB process
Restricted Access: Prepare the customized data
Limit access to only records that need manually review

- Decoupled Data (Kum 2012)
- Automated Honest Broker SW
- Sample selection
- Attribute selection
- Data integration (access to PII)
- Some data cleaning
- Full IRB
- Example: RDC (TX census RDC)
With approved deidentified data
- Locked down VM: customized appliances
- only approved software
- Remote access via VPN
- Very effective for threats from HBC
- Full IRB
- U Chicago-NORC, UNC-Tracs (CTSA), UCSD-iDASH, SAIL

Monitored Access: Freely Repurpose

- Information Accountability model
- Exempt IRB: Explicit data use agreement (5 big Q)
  - Public online (crowdsource)
- Any software & auxiliary data
- Remote Access via VPN
- Less sensitive data (e.g. Aggregate data)
- SHRINE, Secure Unix servers

Open Access: No restriction on use

- Anyone: Publish information for others
- No IRB
- No monitoring use
- Publish data use terms
- Disclosure Limitation Methods (filter)
  - Be careful of incorrect use
- Sanitized data
- Public websites, publications

---

# Privacy Protection Mechanism

<table>
<thead>
<tr>
<th>Access</th>
<th>Restricted Access</th>
<th>Controlled Access</th>
<th>Monitored Access</th>
<th>Open Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protection Approach</strong></td>
<td>Physical restriction to access</td>
<td>Lock down VM (limit what you can do on the system)</td>
<td>Information accountability</td>
<td>Disclosure Limitation</td>
</tr>
<tr>
<td><strong>Monitoring Use</strong></td>
<td>All use on &amp; OFF the computer is monitored</td>
<td>All use on the computer is monitored</td>
<td></td>
<td>Trust</td>
</tr>
<tr>
<td><strong>IRB</strong></td>
<td>Full IRB approved</td>
<td>Full IRB approved</td>
<td>IRB Exempt (register)</td>
<td>Terms of Use</td>
</tr>
<tr>
<td><strong>R1: Cryptographic Attack</strong></td>
<td>Very Low Risk</td>
<td>Low Risk. Would have to break into VM</td>
<td>High Risk</td>
<td>NA</td>
</tr>
<tr>
<td><strong>R2: Data Leakage</strong></td>
<td>Very Low Risk. Memorize data and take out</td>
<td>Physical data leakage (Take a picture of monitor)</td>
<td>Electronically take data off the system</td>
<td></td>
</tr>
</tbody>
</table>
## Comparison of usability

<table>
<thead>
<tr>
<th>U1.1: Software (SW)</th>
<th>Restricted Access</th>
<th>Controlled Access</th>
<th>Monitored Access</th>
<th>Open Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only preinstalled data integration &amp; tabulation SW. No query capacity</td>
<td>Requested and approved statistical software only</td>
<td>Any software</td>
<td>Any software</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>U1.2: Data</th>
<th>Restricted Access</th>
<th>Controlled Access</th>
<th>Monitored Access</th>
<th>Open Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>No outside data allowed But PII data</td>
<td>Only preapproved outside data allowed</td>
<td>Any data</td>
<td>Any data</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>U2: Access</th>
<th>Restricted Access</th>
<th>Controlled Access</th>
<th>Monitored Access</th>
<th>Open Access</th>
</tr>
</thead>
</table>
Deployed together the four data access models can provide a comprehensive system for privacy protection, balancing the risk and usability of secondary data in population informatics research.

Vocab: Information Privacy

- What is *information* privacy?
- Privacy vs confidentiality
  - don’t ask vs don’t tell
- Privacy vs security
- PHI: Protected Health Information
  - Covered entity, covered function
- PII: Personally Identifiable Information
- Coded data
Vocab: Informed Consent

- Opt in
- Opt out
- Blanket consent
- Revised Common Rule: Broad consent
  - Once opt out, must be able to respect.
  - Waiver is not possible
Vocab: Disclosure

- Identity disclosure
- Attribute disclosure
- Harm from disclosure
  - Identity theft: SSN, Name, DOB
  - HIV status
- Group disclosure
- Partial disclosure
- Incremental disclosure
- Minimum necessary standard
  - Cost of implementation?
Information Privacy 101: Point One
Privacy is a BUDGET constrained problem

- Differential privacy literature proves each query leads to some privacy loss while providing some utility in terms of data analysis.
- Current protection mechanism in database research is not effective:
  - de-identified data cannot be linked
  - Not sharing enough details: leads to bias, and invalid results
- The goal is to achieve the maximum utility under a fixed privacy budget.
Information Privacy 101: Point two
Information Accountability (Transparency) Works

- **Secrecy**: Hiding information does not support legitimate use
  - In reality, has limited power to protect privacy
  - Severe Consequences related to
    - Accuracy of data and decisions, use of data for legitimate reasons, transparency & democracy

- **Information Accountability support effective use (Credit Report)**
  - Very clear transparency in the use of the data
  - Disclosure: Declared in writing, so when something goes wrong the right people are held accountable (data use agreements)
  - IT WORKS! Primary method used to protect financial data
  - Internet: crowdsourced auditing (public access IRB)
  - Logs & audits: what to log, how to keep tamperproof log

- D.J. Weitzner et al., Information Accountability, Comm. ACM, vol. 51, no. 6, 2008, pp. 82-87.
Information Privacy 101: Point three
Privacy is contextual

- Helen Nissenbaum (NYU Law School): contextual integrity
- A conceptual framework for understanding privacy expectations and their implications developed in the literature on law, public policy, and political philosophy

Privacy Protection / Violation
- Social norms of expectation (on use, sharing etc)
- Due diligence
- Quantifying harm: loss of job
Information Privacy: Myths and fallacies

- “There is no silver bullet to privacy preserving computation”

- Manage risk by knowing how to handle the tools appropriately

- Privacy by Design: a well orchestrated system to enhance privacy
  - Good IRB approval guidelines
  - Well designed systems to conduct analysis
    - Minimum necessary
    - Fine grained access control
  - Education & Training
  - Regular privacy audits
Thank you

- Secure access models available at TAMU: Virtual Data Library (ViDaL)
  - https://vidal.tamu.edu/
- Questions?
  - Hye-chung Kum, kum@tamu.edu
- Population Informatics Lab
  - https://pinformatics.org/