Privacy and Access

Micah Altman
Director of Research--MIT Libraries
DISCLAIMER

These opinions are my own, they are not the opinions of MIT, any of the project funders, nor (with the exception of co-authored previously published work) my collaborators.

Secondary disclaimer:

“It’s tough to make predictions, especially about the future!”

--- Attributed to Woody Allen, Yogi Berra, Niels Bohr, Vint Cerf, Winston Churchill, Confucius, Disraeli [sic], Freeman Dyson, Cecil B. Demille, Albert Einstein, Enrico Fermi, Edgar R. Fiedler, Bob Fourer, Sam Goldwyn, Allan Lamport, Groucho Marx, Dan Quayle, George Bernard Shaw, Casey Stengel, Will Rogers, M. Taub, Mark Twain, Kerr L. White, etc.
Collaborators & Co-Conspirators

- Privacy Tools for Sharing Research Data Team (Salil Vadhan, P.I.)
  [http://privacytools.seas.harvard.edu/people](http://privacytools.seas.harvard.edu/people)

- Census-MIT Big Data Workshop Series
  - [http://projects.informatics.mit.edu/bigdataworkshops](http://projects.informatics.mit.edu/bigdataworkshops)

- Research Support
  - Supported in part by NSF grant CNS-123723
  - Supported in part by the Sloan Foundation
  - Supported in part by the U.S. Census Bureau
Related Work

Main Project:

- Privacy Tools for Sharing Research Data
  http://privacytools.seas.harvard.edu/

Related publications:


Slides and reprints available from:
informatics.mit.edu

Privacy and Access
Today’s Perspectives & Provocations

* Privacy *
* Utility (Theory) *
* Use (Practice) *
* Calibrating Access and Privacy *
Privacy
Some Definitions

Privacy
Control over extent and circumstances of sharing

Confidentiality
Control over disclosure of information

Identifiability
Potential for learning about individuals based on their inclusion in a data

Sensitivity
Potential for harm if information disclosed and used to learn about individuals
Trend: Privacy in the Spotlight

- Recognition that large amounts of information about people is publicly available
- Breaches and misuse of information is publicized widely – frequent media attention
- Individuals are increasingly concerned about treatment of their information across sectors
- Vast majority of consumers agree consumers that they have lost control of how personal information is collected and used by companies.
  - Many individuals do not understand how and what data is shared

More Information
- Big Data And Privacy: A Technical Perspective, PCAST, 2016. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf)
Privacy is not Anonymization

- Anonymization / deidentification / PII are legal concepts
  - May be no rigorous formal definition
- Definition varies by law, may include ...
  - Presence of specific attributes (e.g., PII, HIPAA identifiers)
  - Feasibility of record linkage ...
  - Evaluation of knowledge of data publisher (e.g. “no actual knowledge”, “readily ascertainable”)
- Legal definitions do not match statistical/scientific concepts
Scientific Trend: More Sophisticated “Reidentification” Concepts

Record-linkage “where’s waldo”
- Match a real person to precise record in a database
- Examples: direct identifiers.
- Caveats: Satisfies compliance for specific laws, but not generally; substantial potential for harm remains

Indistinguishability “hidden in the crowd”
- Individuals can be linked only to a cluster of records (of known size)
- Examples: K-anonymity, attribute disclosure
- Caveats: Potential for substantial harms may remain, must specify what external information is observable, & need diversity for sensitive attributes

Limited Adversarial Learning “confidentiality guaranteed”
- Formally bounds the total learning about any individual that occurs from a data release
- Examples: differential privacy, zero-knowledge proofs
- Caveats: Challenging to implement, often requires interactive systems

More Information
Privacy isn’t always

The “Doesn’t Stay in Vegas” problem -- information shared locally can be found anywhere

- Digitization and API’s make aggregation easy
- Many “public records” were not practically public – until now
- Creates new business models
- Provokes sector-by-sector legislative reaction

More Information
Privacy controls should aim to **Prevent Harm from Disclosure**

- **Data Subjects**
- **Vulnerable Groups**
- **Institutions**
- **Society**
Utility (Theory)
What use is it?

- **Utility** is defined broadly as the analytical value of the data
- There is no universal operational definition of utility
Some Approaches to Measuring of Utility

Statistical Inference
- Precision
- Bias
- Variance

Semantic
- Truthfulness
- Completeness
- Consistency
- Interpretability

Computational
- Entropy
- Information Complexity

$\mathbf{x} \in D \rightarrow \mathbf{x} \in D'$

Image Sources
- [https://commons.wikimedia.org/wiki/File:Solna_Brick_wall_Strecher_bond_variation1.jpg](https://commons.wikimedia.org/wiki/File:Solna_Brick_wall_Strecher_bond_variation1.jpg)
- [https://commons.wikimedia.org/wiki/File:Archery_Target_80cm.svg](https://commons.wikimedia.org/wiki/File:Archery_Target_80cm.svg)
No-Free-Lunch for Privacy & Utility

Any data analysis that is useful* leaks some measurable private information

(However, new methods can sometimes do better than traditional anonymization on both fronts.)
Use (Practice)
Challenges to Measuring Utility

- Utility may not track quality, or value
  - Quality depends on intended use
  - Value is dependent on intended and unintended uses; and on markets

- Utility at one level doesn’t match utility at another levels

Truthful records ≠ accurate aggregates
# Forms of Data Release

## Published Estimates
- Official Indicators
- Pre-computed published tables

## Quick Lookups
- Interactive queries to find a single number or table
- Based on pre-computed tables

## Dynamic Tables & Maps
- Public interactive servers
- Based on public use tabulations or micro-data

## Public Use Tabulations
- Aggregated to pre-defined geographical or logical units
- Processed statistical disclosure limitation methods
- Based on protected micro-data

## Public Use Micro-data
- Processed with SDL: deidentification, sampling, synthetic data
  - In rare cases, synthetic data used
  - Based on protected micro-data

## Protected Micro-data
- Possibly identified
- Available within Research Data Centers
Special Use Challenges

- Computational Replication / Process Verification / Historical Verification
- Data Integration across Separate Source
- Reuse, Reanalysis, Extension

Privacy and Access
Data Reuse

Data Integration
- Pooling Measures (Vertical Partitioning)
- Pooling Subjects (Horizontal Partitioning)
- Pooling Studies (Meta-analysis)
- Pooling Time (Longitudinal Integration)
- Follow-ups/recontacts (Adding Waves, Future data collection)

Data Reuse
- Research Extension / Reanalysis / Critique
- New Model / Analysis / Estimation
- Extraction of New Signals (Measures)
Data Integration and Privacy Challenges

**Aim:**
*Integrate evidence across multiple independent databases*

- Horizontal integration possible – if known to be different sample observations of same population
- Most privacy protections impede
  - Vertical (multi-measure)
  - Longitudinal integration
  - Full meta-analysis
- Anonymization impedes follow-up studies
- Mitigating Approaches
  - Tiered access
  - Third-party escrow
  - Artificial linking identifiers
    - Can leak information
    - Not robust to errors in identifying information
    - Challenging to coordinate across institutions
  - Secure multiparty computation
Aim: Critique previous results, extend model/methods, ask new questions – possibly far in future

- Methods that impede data integration partially impede critique, extension, new questions
- Synthetic data approaches, suppressing measures, can impeded new questions
- Privacy protections that affect interpretability are a challenge to meaningful long-term access
- Generalizations, synthetic data approaches, aggregation impeded extraction of new measures for new questions

Mitigating Approaches
- Broad consent
- Tiered access
Replication/Verification

- Some common characteristics
  - Uses the “original” data, model, methods, implementation
  - Corresponds to those used when results were published
  - Often archived as “replication packages”

- Calibration
  - Verify understanding of software, method, model before extending
  - Often a preliminary to reuse

- Check research integrity
  - Recreate, as closely as possible, the process actually followed by an author in producing previously published results
  - Detect substantial omissions in methods, Unsophisticated data manipulation, Misleading precision in publication

- Retrospective process quality evaluation
  - Evaluate process methods in practice
  - Detect: Transcription errors, Numerical issues, Software bugs

- Historical/legal verification
  - Verify some phenomenon/relation existed/did not exist in data?
  - Verify historical aggregate
  - Did some individual have/not have a particular property

Privacy and Access
Replication, Verification and Privacy Challenges

- Verify particular process and results from a prior publication/statement
  - Affected by variance, bias, interpretability, completeness
- Calibrate prior to extension, critique
  - Affected by bias, variance, interpretability
- Assess process/method robustness
  - Affected by precision, variance, bias
- Verify historical statements about records
  - Affected by truthfulness, completeness

Mitigating Approaches

- Tiered access
- Consent for replication
- Privacy aware publishing
  - Publish research only on post-protected data
  - Avoid publishing results with unnecessary precision
  - Document uncertainty in
- Replication servers
- Process logs + audit trails
Big Data Research and Privacy Challenges

- **Big Data can be Rich, Messy & Surprising**
  - The “Blog problem”:
    Pseudonymous communication used for topic mining can be linked through stylometric analysis

- **Observable Behavior Leaves Unique “Fingerprints”**
  - The “GIS” problem:
    Location trails are individualistic, externally observable, difficult to mas

- **Traditional Anonymization Methods can destroy utility**
  - The “Netflix Problem”:
    many people may have unique long-tail behavior

- **Unintended Algorithmic Discrimination**
  - Big data algorithms may pick up unanticipated relationships in data
  - Algorithms that incorporate human behavior may amplify human biases
  - Can’t prevent by excluding sensitive characteristics as input

More Information

Computational Methods Beyond Anonymization

- **Controlling Access**
  - Virtual Data Enclaves

- **Controlling Computation**
  - Secure Multiparty Computation
  - Functional Encryption
  - Homomorphic Encryption
  - Blockchain

- **Controlling Inference**
  - Differential Privacy

- **Restricting Use**
  - Executable Policy Languages

More Information
- Altman M, Capps C, Prevost R. Location Confidentiality and Official Surveys. Social Science Research Network [Internet]. 2016

Privacy and Access
Opportunities for Public-Private Partnership

- Target gaps between state-of-the-art & practice
- Identify and characterize key public access use cases
- Integrate computational privacy controls
- Systematically analyze and predict informational harms
  - Institutional harm
  - Vulnerable populations
  - Group privacy

Relies on developing ongoing public-private research and development partnerships
Calibrating Privacy and Access
# Catalog of privacy controls

- Procedural, technical, educational, economic, and legal means for enhancing privacy—at each stage of the information lifecycle

<table>
<thead>
<tr>
<th>Access/Release</th>
<th>Procedural</th>
<th>Economic</th>
<th>Educational</th>
<th>Legal</th>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Access controls; Consent; Expert panels; Individual privacy settings; Presumption of openness vs. privacy; Purpose specification; Registration; Restrictions on use by data controller; Risk assessments</td>
<td>Access/Use fees (for data controller or subjects); Property rights assignment</td>
<td>Data asset registers; Notice; Transparency</td>
<td>Integrity and accuracy requirements; Data use agreements (contract with data recipient)/ Terms of service</td>
<td>Authentication; Computable policy; Differential privacy; Encryption (incl. Functional; Homomorphic); Interactive query systems; Secure multiparty computation</td>
</tr>
</tbody>
</table>
Calibrating Controls

Illustrating how to choose privacy controls that are consistent with the uses, threats, and vulnerabilities at each lifecycle stage.

More Information

Calibrating privacy and security controls to the intended uses and privacy risks associated with the data;

When conceptualizing informational risks, considering not just reidentification risks but also inference risks, or the potential for others to learn about individuals from the inclusion of their information in the data;

Addressing informational risks using a combination of privacy and security controls rather than relying on a single control such as consent or deidentification;

Anticipating, regulating, monitoring, and reviewing interactions with data across all stages of the lifecycle (including the post-access stages), as risks and methods will evolve over time; and
Recommended Readings