# CCC Engineering Privacy Workshop 2015

# Practical Deidentification

## Anonymization and Risk

Ira Rubinstein
NYU School of Law

# Background

- Rubinstein & Hartzog, Anonymization and Risk (forthcoming Washington Law Review, 2016)
- Anonymization has spawned a large literature but critics and defenders of deidentification techniques remain sharply divided
- The policy debate has stagnated
- Topic needs reframing *away* from the quest for perfect anonymization and *towards* describing a process of risk management

# Main thesis

- Best way to move the deidentification debate past the alleged failures of anonymization is to reframe the debate in terms of data release policy and focus on the process of minimizing risk, not preventing harm.
  - (In other words, think of data release more in terms of data security policy)

# A Few Corollaries

1. “Data release” policy suggests a range of methods and techniques
   - Direct access
     - Licensing and security enclaves
   - Dissemination-based access
     - Deidentification
   - Query-based access
     - Differential privacy

# A Few Corollaries (continued)

- Use the full spectrum of data release protections tailored to an organization's anticipated risk.
- This implies combining the most appropriate technical methods with available legal tools:
  - Informed consent, tiered access, security obligations, data use agreements, statutory prohibitions on reuse, transfer, reidentification, etc.
- [Query: What is the proper relationship between engineering solutions (especially PETs) and various legal mechanisms?]
  - E.g.: Deidentification + rule on open vs. controlled access

# A Few Corollaries (continued)

- The law of data release should look more like the law of data security:
  - Process-based,
  - Contextual, and
  - Tolerant of harm,
- Provided that procedures to minimize risk are implemented *ex ante*.

# How We Got Here

- Deidentification debates fixated on a few high profile incidents
- Legal scholars have joined the fray by adopting extreme positions
- Technical community equally divided
  - Not only in how they view the implications of the auxiliary information problem, but in their goals, methods, interests, and measures of success.
  - For better or worse, we adopt typology of "pragmatists" and "formalists"
  - Mistake to fixate on a single approach

# Risk Factors

- Ours is a risk-based approach. So what are the risk factors?
  - Data volume
  - Data sensitivity
  - Type of data recipient
  - Data use
  - Data treatment technique
  - Data access controls
  - Consent
  - Consumer expectations

# Legal Reforms

- Develop a reasonableness standard for data release, administered by federal agencies including FTC:
  1) Assess data to be shared and risk of disclosure;
  2) Minimize data to be released;
  3) Implement reasonable data control techniques as appropriate;
  4) Develop a monitoring, accountability, and breach response plan.
- Other reforms
  - Address "broken promises" of anonymization (i.e., deceptive claims)
  - Revise the definition of PII/personal data
  - Revise the HIPAA safe harbor

# A Path Forward

- Salil Vadhan and his colleagues have proposed that regulatory agencies maintain a safe harbor list of data-sharing mechanisms appropriate for different contexts that can be maintained and regularly updated with the input of experts and stakeholders.

# More on "Safe-Harbor Lists"

- Each entry in this list would specify:
  - A class of data sources (e.g., electronic health records vs. genomic data)
  - A class of data-sharing methods (e.g. HIPAA-style deidentification or differential privacy)
  - A class of informed consent mechanisms
  - A class of potential recipients
- A safe harbor approach therefore requires a mapping exercise undertaken by technical and legal experts and other interested stakeholders