Practical Approaches to Big Data Privacy Over Time

Alexandra Wood

Fellow, Berkman Klein Center for Internet & Society

with Micah Altman, David O’Brien, & Urs Gasser

Presentation for the Brussels Privacy Symposium
November 8, 2016

Corporations and governments are collecting data more frequently, and collecting, storing, and using it for longer periods.

Greater quantities of personal data are being collected, at finer levels of granularity, at more frequent intervals.

These activities create opportunities for research, but also increase identifiability and expand the range of harms to which individuals are exposed.

Current accepted practices for protecting privacy in long-term data are highly varied across research, commercial, and government contexts.

Businesses and governments generally rely on approaches such as the exclusive use of notice and consent mechanisms and de-identification techniques.

These practices differ substantially from the frameworks and interventions used by researchers and institutional review boards to address challenges associated with managing privacy in long-term research data activities.

Research settings: Privacy practices in long-term research studies are

- governed by strict ethical and legal frameworks, including oversight by an IRB,
- heavily curated by their investigators, and
- incorporate multiple layers of protection, including explicit consent, systematic design and review, statistical disclosure control, and legal/procedural controls.

Commercial and government settings: Industry and government actors generally

- operate within a legal framework that has arguably been slower to evolve to address data privacy and ethical challenges,
- rarely engage in systematic review of privacy risks and planning for long-term review, storage, use, and disclosure,
- rely on a narrower subset of privacy controls such as notice and consent and de-identification.

The expanding timescale and new commercial uses are increasing risks and decreasing the effectiveness of current approaches.

The increasing scale of commercial and government data programs is putting pressure on current privacy practices, due in part to the following factors:

- The collection of data at more frequent intervals,
- The extended period of data collection, and
- The amount of time that has elapsed between collection and use.

The age of the data, or the duration of storage and use of personal data, alters privacy risks.

The effect of age on risk is complex.

Associated with a decrease in risk: Observable characteristics generally change over time. Availability and accuracy of data have historically decreased with time. Individuals may be less vulnerable to harm from older data.

Associated with an increase in risk: As data are digitized, more widely disseminated, and made persistently available, risks increase. Data are stored for longer periods of time, increasing the likelihood of data breaches. Threats from data use increase, as data are more likely to be used in ways that were unanticipated when collected.

Longer periods of data collection create additional privacy risks.

Data covering a longer period, i.e., data that describe trends, may result in increased threats.

Example: Data collected over the course of a long-term medical research study can reveal information about an individual’s development of risk factors for, or progression of, heart disease, diabetes, and Alzheimer’s disease or dementia.

Longer periods may also be correlated with greater age of the data, and the interaction with high frequency may enable increased detection of trends, further increasing threats and enabling stronger identification of unique patterns of behavior.

High-frequency data pose a significant challenge to traditional privacy approaches.

Commercial and government big data collection can lead to more frequent observations of individual behavior.

For example, the microphone, camera, accelerometer, GPS receiver, and other sensors embedded in a mobile device can generate fine-grained data.

Continuous monitoring of such observations can reveal sensitive facts about an individual’s health and behavior.

High-frequency data also dramatically increase identifiability by revealing unique patterns of behavior.

Table 1. Key risk drivers for big data over time and their effects on privacy risk components.

	Identifiability	Threats (sensitivity)	Vulnerabilities (sensitivity)
Age	Small decrease	Moderate increase	Moderate decrease
Period	Small increase	Moderate increase	No substantial evidence of effect
Frequency	Large increase	Small increase	No substantial evidence of effect

Non-temporal risk factors of big data also affect privacy risk components in different ways.

High-dimensional data pose challenges for traditional privacy approaches such as de-identification, and can support new uses of data that were unforeseen at the time of collection.

Broader analytic uses, such as the use of data for personalized classification, and both traditional and modern approaches to de-identification fail to protect against learning facts about populations that could be used to discriminate.

Increases in sample size and diversity lead to heightened risks that a target individual is included, vulnerable populations are included, and a wide range of threats are plausible.

Table 2. Key non-temporal risk drivers for big data and their effects on risk.

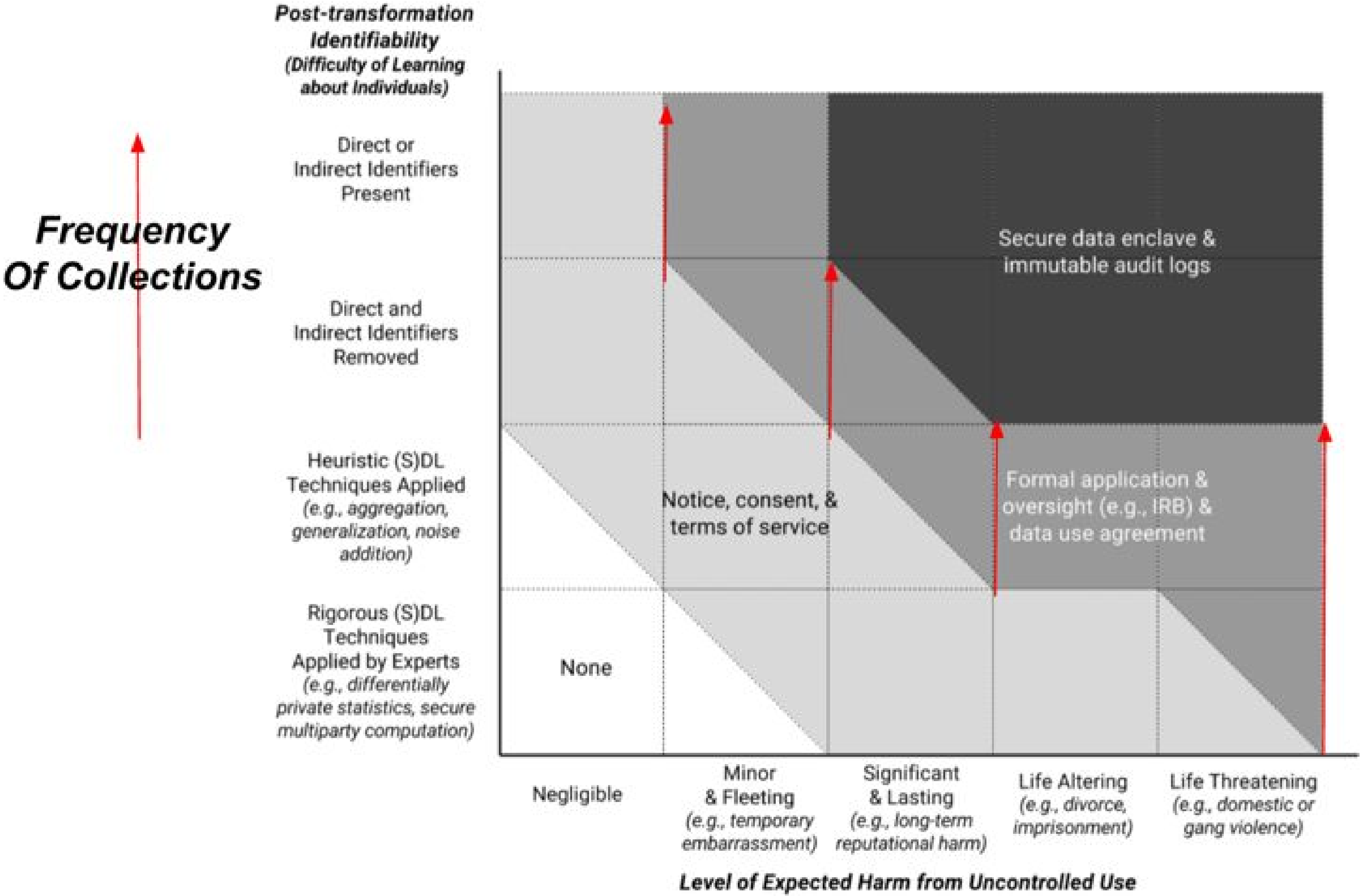
	Identifiability	Threats (sensitivity)	Vulnerabilities (sensitivity)
Dimensionality	Moderate increase	Moderate increase	No substantial evidence of effect
Broader analytic use	Large increase	Moderate increase	Large increase
Sample size	Small increase	No substantial evidence of effect	Moderate increase
Population diversity	Small decrease	Moderate increase	Small increase

Long-term data risk factors change the surface of suitable privacy controls.

The risk-benefit analyses and best practices established by the research community can be instructive for privacy management in other settings.

Appropriate solutions can be informed by analyzing the relationship between identifiability, sensitivity, and the suitability of various procedural, legal, and technical controls used in long-term research.

Practical data sharing models can combine different types of interventions for evaluating and mitigating risk, balancing privacy and utility, and providing enhanced transparency, review, and accountability.


Post-transformation Identifiability
(Difficulty of Learning about Individuals)
Frequency Of Collections
Direct or Indirect Identifiers Present
Direct and Indirect Identifiers Removed
Heuristic (S)DL Techniques Applied (e.g., aggregation, generalization, noise addition)
Rigorous (S)DL Techniques Applied by Experts (e.g., differentially private statistics, secure multiparty computation)
Secure data enclave & immutable audit logs
Notice, consent, & terms of service
Formal application & oversight (e.g., IRB) & data use agreement
None
Negligible
Minor & Fleeting (e.g., temporary embarrassment)
Significant & Lasting (e.g., long-term reputational harm)
Life Altering (e.g., divorce, imprisonment)
Life Threatening (e.g., domestic or gang violence)
Level of Expected Harm from Uncontrolled Use

Selection from the wide array of privacy controls

Identifiability-focused controls:

- Simple redaction,
- Heuristic disclosure limitation techniques, and
- Robust disclosure limitation (e.g., secure multiparty computation and differentially private statistics)

Combinations of sensitivity-focused controls:

- Secure data enclaves with auditing procedures (and, in some cases, secure multiparty computation, computable policies, or personal data stores),
- Formal application and review by an ethics board, and
- Notice, consent, and terms of service (and personal data stores, blockchain tools, and privacy icons)

Thank you

Research collaborators: Privacy Tools for Sharing Research Data project (http://privacytools.seas.harvard.edu)

Research support from the National Science Foundation and the Alfred P. Sloan Foundation

FUTURE OF
PRIVACY
FORUM

VUB

BRUSSELS
PRIVACY
HUB