Practical Approaches to Big Data Privacy Over Time

Micah Altman,[1] Alexandra Wood,[2] David R. O'Brien[3] & Urs Gasser[4]

Abstract prepared for submission to the 2016 Brussels Privacy Symposium

Corporations and governments are collecting, storing, analyzing, and sharing detailed information about individuals over increasingly long periods of time. Vast quantities of data from new sources and novel methods for large-scale data analysis promise to yield a deeper understanding of human characteristics and behavior and, in turn, to advance science, public policy, and innovation. At the same time, the collection and use of fine-grained personal data over time is understood to be associated with significant and growing risks to individuals, groups, and society at large.

Although commercial firms and government agencies have implemented measures to address data privacy risks, approaches in widespread use represent a limited subset of the privacy protection techniques that are available. For instance, it is a common practice when collecting, storing, and sharing data about individuals to protect privacy by de-identifying data through the removal of personally identifiable information. However, there is growing evidence that, while it reduces some risks, de-identification alone does not protect information in the manner that most individuals expect and often results in unnecessary removal of useful information. Additionally, the expanding scale of big data collection and long-term storage is leading to massive accumulations of personal data that threaten to further erode the effectiveness of traditional de-identification techniques. In light of these challenges for de-identification in the era of big data, scholars and practitioners are now exploring new technical, procedural, and legal interventions for managing data privacy that can complement traditional approaches.

In this paper, we compare real-world data collection and management programs across government, industry, and research settings. This examination focuses on identifying the characteristics that drive the risks and benefits of these programs, as well as the specific technical, procedural, contractual, and regulatory controls in use. We observe that current debates around privacy in commercial big data and open government data are reminiscent of earlier debates regarding the long-term risks associated with human subjects research data. Data collected throughout the course of a longitudinal research study are in many cases highly specific, identifiable, and sensitive, and carry risks that are similar to those associated with personal data held by corporations and governments. For decades, researchers and institutional review boards have intensively studied the long-term risks in this context, and have developed practices that address many of the challenges associated with obtaining informed consent and de-identifying data. For these and other reasons, the risk-benefit analyses and best practices established by the research community can be instructive for privacy management with respect to the long-term collection and use of sensitive commercial and government data.

This paper presents lessons that can be learned from longstanding research data management practices

[1] MIT Libraries, Massachusetts Institute for Technology, escience@mit.edu

[2] Berkman Klein Center for Internet & Society, Harvard University, awood@cyber.law.harvard.edu

[3] Berkman Klein Center for Internet & Society, Harvard University, dobrien@cyber.law.harvard.edu

[4] Berkman Klein Center for Internet & Society, Harvard University, ugasser@cyber.law.harvard.edu

and potentially applied in newly emerging commercial big data and open government data programs. Different legal frameworks and institutional constraints often apply and lead to variations in practice across these settings. However, there are notable similarities in the data characteristics, privacy risks, and challenges involved, and it may be appropriate to apply principles for balancing privacy and utility in data releases more universally. For example, corporations and governments may consider adopting review processes similar to those that have been established at research institutions to systematically analyze the risks and benefits associated with data collection, retention, use, and disclosure over time. Rather than relying on a single intervention such as de-identification or consent, corporate and government actors may weigh the suitability of combinations of interventions from the wide range of privacy and security controls that are available. In particular, new procedural, legal, and technical tools for evaluating and mitigating risk, balancing privacy and utility, and providing enhanced transparency, review, accountability, are being investigated to be deployed as part of comprehensive research data management plans. The suitability of new tools, especially formal privacy models such as differential privacy, should likewise be explored within the context of corporate and government settings. We conclude with practical recommendations for calibrating combinations of privacy and security controls, including notice and consent, de-identification, ethical review processes, differential privacy, and secure data enclaves, to the intended uses and privacy risks associated with a specific corporate or government data program.