Part I Conceptual Framework

This part begins by considering the existing legal constraints on the collection and use of big data in the privacy and confidentiality context. It then identifies gaps in the current legal landscape and issues in designing a coherent set of policies that both protect privacy and yet permit the potential benefits that come with big data. Three themes emerge: that the concepts used in the larger discussion of privacy and big data require updating; that how we understand and assess harms from privacy violations needs updating; and that we must rethink established approaches to managing privacy in the big data context.

The notion of 'big data' is interpreted as a change in paradigm, rather than solely a change in technology. This illustrates the first central theme of this part of the book. Barocas and Nissenbaum define big data as a "paradigm, rather than a particular technology," while Strandburg differentiates between collections of data, and collections of data that have been "datafied," that is, "aggregated in a computationally manipulable format." She claims that such datafication is a key step in heightening privacy concerns and creating a greater need for a coherent regulatory structure for data acquisition. Traditional regulatory tools for managing privacy – notice and consent – have failed to provide a viable market mechanism allowing a form of self-regulation governing industry data collection. Strandburg elucidates the current legal restrictions and guidance on data collection in the industrial setting, including the Fair Information Practice Principles (FIPPs) dating from 1973 and underlying the Fair Credit Reporting Act (FCRA) from 1970 and the Privacy Act from 1974. Strandburg advocates a more nuanced assessment of trade-offs in the big data context, moving away from individualized assessments of the costs of privacy violations. The privacy law governing the collection of private data for monitoring purposes should be strengthened, in particular, a substantive distinction should be made between datafication and the repurposing of data that was collected as a byproduct of providing services. Additionally, she suggests taking a substantive approach to the ideas of notice and consent in particular to clarify their meaning for large entities.

The inadequacy of assessment of harm from privacy breaches is another major theme of this part, extending it from the level of the individual to that of groups or classes, and even society as a whole. Acquisti notes that this requires a greater understanding of the breadth of types of privacy breaches and of the nature of harm as diffused over time, and an improved valuation of privacy in the big data context. Consumers may value their own privacy in variously flawed ways. They may have incomplete information, for example, or be given an overabundance of information that renders processing impossible, or use heuristics that systematize deviations from rational decision making. Acquisti notes that privacy protection can both potentially increase and decrease economic efficiency in the marketplace, and that deriving benefits from big data may not conflict with benefits from assuring privacy protection.

In order to address these issues, several authors ask us to rethink traditional approaches to privacy. This is the third overarching theme of this part of the book. Barocas and Nissenbaum argue that the concepts of anonymity and informed consent do not create solutions to the privacy issue. As datasets become increasingly linked, anonymity is largely impossible to guarantee in the future. This also implies that it is impossible to give truly informed consent, since we cannot, by definition, know what the risks are from revealing personal data either for individuals or for society as a whole.

The use of privately collected data is largely unregulated. Ohm describes the few regulations that do apply – such as the Health Insurance Portability and Accountability Act (HIPAA), the Privacy Act, and the Fair Credit Reporting Act (FCRA) – and explains that the United States employs a 'sectoral' approach to privacy regulation, in that different economic areas have separate privacy laws. Ohm also calls into question the traditional notion of notice in the case of big data. To whom are you to give notice, and for what? The results of big data analysis can be unpredictable and sometimes unexplainable, another reason it is difficult to assess privacy risks accurately in the big data context.

Ohm advocates a new conceptualization of legal policy regarding privacy in the big data context, guided by five principles for reform: (1) rules must take into account the varying levels of inherent risk to individuals across different datasets, (2) traditional definitions of personally identifiable information need to be rethought, (3) regulation has a role in creating and policing walls between datasets, (4) those analyzing big data must be reminded, with the frequency in proportion to the sensitivity of the data, that they are dealing with people, and (5) the ethics of big data research must be an open topic for continual reassessment.

In the final chapter, Stodden focuses on the theme of research integrity in the big data context. She notes a conflict between research requirements regarding the replication of computational results, which can require data access, and traditional methods of privacy protection via sequestration. She advocates establishing 'middle ground' solutions whenever possible that maximize verification of computational findings, while taking into account any legal and ethical barriers. Permitting authorized researchers access to confidential data within a 'walled garden' can increase the ability of others to independently replicate big data findings, for example. Two principles are presented to help guide thinking regarding reproducibility and verification in big data research: the Principle of Scientific Licensing and the Principle of Scientific Data and Code Sharing. That is, in the scientific context, legal encumbrances to data sharing for purposes of independent verification should be minimized wherever possible, and access to the data and methods associated with published findings should be maximized subject to legal and ethical restrictions.

1 Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context

Katherine J. Strandburg

Introduction

Knowledge is power. 'Big data' has great potential to benefit society. At the same time, its availability creates significant potential for mistaken, misguided, or malevolent uses of personal information. The conundrum for law is to provide space for big data to fulfill its potential for societal benefit, while protecting citizens adequately from related individual and social harms. Current privacy law evolved to address different concerns and must be adapted to confront big data's challenges. This chapter addresses only one aspect of privacy law: the regulation of private sector acquisition, aggregation, and transfer of personal information.[1] It provides an overview and taxonomy of current law, highlighting the mismatch between current law and the big data context, with the goal of informing the debate about how to bring big data practice and privacy regulation into optimal harmony.

Part I briefly describes how privacy regulation in the United States has evolved in response to a changing technological and social milieu. Part II introduces a taxonomy of privacy laws relating to data acquisition, based on the following features: (1) whether the law provides a rule- or a fact-based standard; (2) whether the law is substantive or procedural, in a sense defined below; and (3) which mode(s) of data acquisition are covered by the law. It also argues that the recording, aggregation, and organization of information into a form that can be used for data mining, here dubbed 'datafication', has distinct privacy implications that often go unrecognized by current law. Part III provides a selective overview of relevant privacy laws in light of that taxonomy. Section A discusses the most standards-like legal regimes, such as the privacy torts, for which determining liability generally involves a fact-specific analysis of the behavior of both data subjects and those who acquire or transfer the data ('data handlers'). Section B discusses the Federal Trade Commission's (FTC's) 'unfair and deceptive trade practices' standard,[2] which depends on a fact-specific inquiry into the behavior of

data handlers, but makes general assumptions about data subjects. Section C discusses rule-like regimes, such as the Privacy Rule[3] of the Health Insurance Portability and Accountability Act[4] (HIPAA Rule). Part IV points out some particular features of the mismatch between current law's conceptualization of the issues and the big data context, using the taxonomy developed in Part II as an aid to the analysis. It then makes several suggestions about how to devise a better fit.

I. The Evolution of U.S. Privacy Law

Outside of the law enforcement context, privacy law was erected on the foundation of Warren and Brandeis's famous 1890 article, *The Right to Privacy*.[5] The privacy torts built on that foundation were concerned primarily with individualized harms of emotional distress, embarrassment, and humiliation arising out of 'intrusion upon seclusion' or 'public disclosure of private facts'. Privacy law also aimed to protect confidentiality in certain kinds of relationships, often involving professional expertise, in which information asymmetry and power imbalances create a potential for exploitation. These torts provide compensation for individualized injuries caused by egregious deviations from social norms. In principle, and often in fact, the tort paradigm employs a highly contextualized analysis of the actions of plaintiff and defendant and the relationship between them.

In the 1970s, the development of digital computers and the increasing complexity of the administrative state led to an expansion in 'computer-based record-keeping operations' by governments and certain other large institutions, such as banks. This expansion raised fears of misuse, unfairness, lack of transparency in decision making, and chilling of autonomous behavior distinct from the concerns about emotional distress and reputation at the heart of the privacy torts. Fair Information Practice Principles (FIPPs), which have become the mainstay of data privacy law, were developed during this period as an approach to those issues. The Fair Credit Reporting Act (FCRA),[6] adopted in 1970, and the Privacy Act of 1974,[7] regulating data use by government agencies, were based on FIPPs.

A set of five FIPPs were proposed in 1973 in a report commissioned by the Department of Health, Education, and Welfare (HEW Report):

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.

3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.[8]

These principles, along with three sets of related 'safeguard requirements',[9] attempted to cope with the scale of data collection by substituting transparency and consent for the individualized fact-specific approach of the privacy torts. The HEW Report recognized the difficulty of legislating substantive rules in light of the "enormous number and variety of institutions dealing with personal data," arguing that institutions should be "deterred from inappropriate practices, rather than being forced by regulation to adopt specific practices."[10]

Another important version of FIPPs was formulated by the OECD in 1980 (OECD FIPPs), articulating eight principles, which expanded on the HEW FIPPs and include a "Collection Limitation Principle" that there "should be limits to the collection of personal data," a "Data Quality Principle" that data collected should be "relevant to the purposes for which they are to be used" and "accurate, complete and kept up-to-date," and a "Purpose Specification Principle" that purposes should be specified in advance and "subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes."[11]

At the turn of the 21st century, the FTC, which has taken the primary role in commercial privacy regulation, recommended a rather slimmed down set of FIPPs for online privacy:

(1) Notice – Web sites would be required to provide consumers clear and conspicuous notice of their information practices. . . .
(2) Choice – Web sites would be required to offer consumers choices as to how their personal identifying information is used beyond the use for which the information was provided (e.g., to consummate a transaction). . . .
(3) Access – Web sites would be required to offer consumers reasonable access to the information a Web site has collected about them, including a reasonable opportunity to review information and to correct inaccuracies or delete information.

(4) Security – Web sites would be required to take reasonable steps to protect the security of the information they collect from consumers.[12]

In practice, outside of a few sectors, such as health care, the FIPPs approach in the United States has been whittled down to a focus on procedures for ensuring notice and consent. What explains this progression? Cheap and ubiquitous information technology makes large-scale data aggregation and use possible for a wide variety of private sector entities in addition to the government agencies and large institutions, such as banks, that were the subject of privacy concerns in the 1970s and 1980s. The resulting expansion of private sector data collection and use has competing implications for privacy regulation. On the one hand, the acquisition and use of an increasing quantity and scope of personal information by an increasingly large and various set of private entities heightens privacy concerns relating to data security and breach, accountability and transparency, and unpredictable data uses. On the other hand, substantive regulation of a large and diverse array of private sector entities is politically controversial, regulations that effectively span the field of data handlers are hard to devise, and monitoring the data practices of such a large number of players is difficult and expensive.

As a result, the trend was to assume (or at least hope) that notice and consent would provide a market mechanism for encouraging industry self-regulation of data privacy. In light of recent acceleration in data collection and the development of big data approaches to mining aggregated data, it now is widely recognized that the notice and consent paradigm is inadequate to confront the privacy issues posed by the big data explosion.[13] The notice and consent paradigm assumes that citizens are able to assess the potential benefits and costs of data acquisition sufficiently accurately to make informed choices. This assumption was something of a legal fiction when applied to data collected by government agencies and regulated industries in the 1970s. It is most certainly a legal fantasy today, for a variety of reasons including the increasing use of complex and opaque predictive data-mining techniques, the interrelatedness of personal data, and the unpredictability of potential harms from its nearly ubiquitous collection.[14]

II. A Taxonomy of Privacy Laws Relevant to Data Acquisition

As mentioned in the introduction, it is useful to organize this selective overview of U.S. privacy law relating to data acquisition and transfer

according to a taxonomy focusing on three characteristics: (A) whether the law takes a rule-like or fact-specific standards-like approach; (B) whether the law regulates substance or procedure, in a sense defined below; and (C) what modes of data acquisition are covered by the law. While these distinctions are not bright lines, an approximate categorization is useful in analyzing the uncomfortable fit between current privacy law and big data projects.

A. Rules or Standards

Privacy law regimes vary according to the extent to which they impose flexible fact-specific standards or generally applicable rules, but can be divided roughly into three groups. Laws in the first group, such as the torts of intrusion upon seclusion and public disclosure of private facts, assess liability using standards that depend on detailed fact-intensive inquiry into the activities of both subjects and acquirers of personal information. Laws in the second group, such as Section 5 of the FTC Act, employ standards-based assessment of the activities of data holders, while relying on presumptions about data subjects. Laws in the third group, such as the HIPAA Rule, mandate compliance with rules.

Trade-offs between rules and standards are endemic to law.[15] Ex ante, rules provide clearer direction for behavior, while standards provide leeway so behavior can be tailored more optimally to specific contexts. Ex post, rules are cheaper to enforce and leave less room for bias, while standards leave more discretion for crafting context-sensitive and fair outcomes. These tensions are dynamic. Because of these trade-offs, courts and legislatures often attempt to draw at least some bright lines (such as a line between private and public) to guide the application of standards, while rule-like regimes often become complex (or even byzantine) as lawmakers try to anticipate all relevant contingencies.

B. Substance or Procedure

Privacy law also grapples with trade-offs between substantive and procedural regulation. Compliance with substantive regulation, as I use the term here, is determined by asking: Was it legally acceptable for Data Handler A to acquire or transfer this information about Data Subject B in this situation and in this way? Compliance with procedural regulation, on the other hand, is determined by asking: Did Data Handler A follow the appropriate procedures in acquiring or transferring information about Data Subject B?

Though substantive regulation is preferable when goals are well defined and outcomes are observable by enforcement agencies, procedural regulation is advantageous in some situations. Regulated entities may be better situated than lawmakers, by virtue of expertise or superior information, to make substantive determinations, especially when circumstances are expected to evolve, perhaps by technological advance, but data holders may have incentives to use their discretion in socially undesirable ways. Procedural regulation may be used to limit their ability to do so. Substantive outcomes may be difficult to specify or to observe. Procedural regulations may structure behavior so as to make desirable outcomes more likely and may make compliance easier to audit. Procedural regulation also may help to prevent negligence and mistake. Procedural and substantive approaches often are combined in privacy regulation. For example, some laws require different procedures based on substantive distinctions between data-handling entities, types of data, or purposes of data acquisition.

C. Modes of Data Acquisition

There are three basic avenues for acquiring big data: monitoring, acquisition as a byproduct of another activity, and transfer of pre-existing information. Monitoring, as I use the term here, applies broadly to the recording of information in plain view and information acquisition by means such as wiretapping and spyware. Acquisition as a byproduct of another activity is common for service providers such as telecommunications providers, utilities, online websites and apps, search engines, and governments.

However it is acquired, if information is to be used in a big data project it must be recorded, quantified, formatted, and stored digitally to make it usable for computational knowledge discovery. Note that what I will call 'datafication' is distinct from digitization. Cellphone photos are digital, but they are not datafied unless they are aggregated in a computationally manipulable format. Datafication has independent privacy implications because recording and organizing information changes the uses to which it can be put, both for good and for ill.[16] Importantly, because computation methods are continually developed and refined, datafication is likely to open the door to uses that were not feasible (and hence obviously not anticipated) at the time the data was acquired.

To illustrate the role of datafication in monitoring, consider video surveillance. Without cameras, the only record of what happens on city streets is in the minds of the human beings who pass by. Those memories are scattered, degrade quickly, may be inaccurate, and are very costly