Position: Privacy and Policy in Polystores: A Data Management Research Agenda

Joshua A. Kroll[1], Nitin Kohli[1], and Paul Laskowski[1]

U.C. Berkeley School of Information

1 Introduction

Modern data-driven technologies are providing us with new capabilities for working with data across diverse storage architectures and analyzing it in unified frameworks to yield powerful insights. Even as the benefits of these technologies come into focus, they are fueling a corresponding debate over the responsible use of data and the urgency of mitigating privacy risks to individuals. This discussion is happening as new analysis techniques, based on correlating data across sources and types and exploiting statistical structure, have challenged classical approaches to privacy, leading to a need for radical rethinking of the meaning of privacy and its application to federated databases.

New technologies have often triggered concerns about the privacy of individuals, dating back to the advent of publishing photographs in newspapers and even earlier [38]. The origin of modern data protection and privacy laws can be traced as far back as 1972, when the public learned that the U.S. government had been holding secret databases of information about citizens. In a formative report created in response to the episode [36], a study committee at the U.S. Department of Health, Education and Welfare articulated a set of core *Fair Information Practices* (FIPs) that have formed the basis of nearly every subsequent privacy and data protection law [14]. Attempting to mitigate these concerns and to operationalize protections around them has led to a rich field of privacy study across several disciplines.

This paper situates these concerns in light of research on polystores [7] and database federation more generally. While the correlation of data across stores presents significant new privacy risks, it also provides new opportunities for interventions and mitigations. This paper outlines this situation and presents a research agenda for research in data privacy for the database community.

2 Traditional Approaches to Privacy in Databases

The study of privacy in computer science largely emerged from the broader study of security, inheriting a focus on concepts like confidentiality. This section briefly surveys traditional approaches, while Section 3 describes ways in which they are insufficient to the task of protecting privacy.

Access control In databases, privacy-as-confidentiality has often meant assigning permissions and roles to database users [2]. Modern applications of this idea

include *data tagging* systems, which maintain metadata about data items to enable access and query policies [5]. Together, these methods support the important goal of limiting who (or which processes) can access what data items in what contexts at what times. At the same time, they do not directly provide a way to protect the semantics of the underlying data items. Moreover, these methods are not tailored to limit what can be learned from data when access is allowed, for example to protect against misuse by insiders or against authorized-but-privacy-impinging uses.

Query and Result Filtering To complement access control and enable more flexible policies, the database community has invested significant research effort in developing query auditing and result filtering systems [25]. Auditing provides two significant improvements over methods purely based on access control: First, auditing allows for flexible, detailed, and context-sensitive policies, as decisions can be based on arbitrary processing of query text, rather than a model based on rows, columns, and user roles. Second, auditing decision can be made in an online fashion, adapting to the history of prior queries based on estimates of the cumulative privacy risk. However, these systems generally define such risks in the frame of controlling individual pieces of data.

Anonymization Another line of work considers how to render information in databases or outputs *anonymous*, meaning that rows of data cannot be associated to a particular individual. This extends the concept of privacy beyond information flow to the protection of individual *identities*. The question of what constitutes an identity is, however, difficult and its answers are hotly contested. Traditional approaches to anonymization focus on syntactic properties of data sets, such as the property that no query can return fewer than k individuals [34], or that any result set has at least ℓ distinct values for a protected sensitive attribute [22]. Many legal definitions of identity treat data as belonging to one of two categories: "personally identifiable information" (PII) and "not personally identifiable information. This separation is easy to implement in code; however, it elides an important fact (discussed in detail in Section 3: syntactic methods are insufficient to prevent the disclosure of attributes of individuals [27] and so have given way to more information-theoretic protections.

Differential Privacy Over the last 15 years, a formal approach to privacy has emerged from a branch of computer science known as *differential privacy*. Researchers in this field design computer systems that employ randomness or noise to conceal information about individuals. Moreover, a mathematical proof characterizes the privacy loss that can befall any individual in the worst case.

The term differential privacy does not refer to a single technique, but rather a mathematical standard. It guarantees that the behavior of a computer system is very similar whether an individual is included in a database or removed [9,8]. The degree of similarity is controlled by a parameter ϵ, which represents the maximum amount of information that an adversary can learn about an individual by studying the mechanism output. Conceptually, a smaller ϵ provides stronger privacy protection, but there is usually a cost in the form of greater noise and thus

limited utility from queries. Many heuristics have been proposed to determine an appropriate value of ϵ [26,21,16,19].

From a different angle, the parameter ϵ can be viewed as a *measure* of (worst-case) privacy loss for a given algorithm. A useful property of differential privacy is additive composition, meaning that when multiple queries are run in sequence, the individual ϵ's can be added together to yield a bound for the entire process [10]. This allows organizations to manage access using a *privacy budget*. Each time a query is run, the corresponding ϵ is subtracted from the budget. When the budget is exhausted, the data must be permanently retired to ensure the privacy guarantee holds.

3 Polystores as privacy risk

We have outlined a toolkit of existing privacy defenses, including access control, query and result filtering, anonymization, and differential privacy. These are all valuable tools, but each has limitations. In this section, we present three stylized facts to explain why existing approaches to privacy are insufficient to protect individuals, and why database federation only exacerbates the problem. To meet evolving risks to privacy, it will therefore be necessary to look beyond traditional technologies and develop new areas of research.

Statistical inference presents new privacy threats. Syntactic approaches to database anonymization – such as k-anonymity and ℓ-diversity preserve the structure of the data and the relationships between data elements. This structure can often be exploited, along with information outside the database, to "re-identify" or "de-anonymize" the database. Early efforts by Narayanan and Shmatikov demonstrated that this could be achieved at large scale [29]. Subsequent work has demonstrated an almost complete failure of anonymization techniques to protect individuals [27]. A corollary of this is that the legal notion of personally identifiable information (PII) (known in the GDPR as "personal data") should now be considered obsolete, and all data should be treated as if it is "personal" or "identifying" unless there is a strong case that it is not [28].

One reason for this is the ease of determining *functional identifiers* in large, rich data sets, or sets of attributes which are held by only one individual in that collection. Such identifiers "single out" individuals, implying a set of binary queries for which the individual has a unique set of answers [6]. Representationally, such queries could be treated as the bits of a numeric identifier, which would clearly qualify as personal and identifiable data under existing legal regimes (see Section 4) [30]. A curious corollary of this argument is that a 33-bit identifier would suffice to identify all humans on earth uniquely. [1] Thus, any reasonably rich data set should be considered to contain such functional identifiers.

Database federation presents new and heightened risks. A central motivation for the federation of disparate databases is the idea that cross-referencing the data across them will yield new and better insights. It is therefore natural that such

[1] $2^{33} = 8,589,934,592$, compared to a world population of around 7.7 billion.

cross-source correlation may enable the increased extraction of sensitive information about individuals, groups of people, or organizations. In fact, federated databases share several features that contribute to increased privacy risks. First, it is clear that cross-referencing data across many sources abstractly creates data items with more attributes, providing more free parameters for the construction of functional identifiers as described above. Second, it is often the case that these systems are built to enable downstream analytics processing or machine learning pipelines. Such processing naturally seeks to determine how to separate, cluster, and classify data items and so can be thought of as inherently and automatically constructing such identification schemes. Finally, consolidation of data across data sources risks undermining assumptions made during the policy establishment process when sources were not federated. Even properties as straightforward as access control or data permissions can be so subverted as to be totally violated when such assumptions change [2]. Relatedly, centralization of data access creates new and attractive points of attack for outside hackers or malicious insiders. Such "attacks" on privacy can also occur without malice, simply through the authorized realization that new insights are possible through analysis of newly combined data, undermining existing business logic.

Differential privacy requires design-level intervention. One might hope that we can take an arbitrary algorithm and make it differentially private. Unfortunately, this is not the case in general. As differentially private algorithms require noise to protect any possible input data value, sufficient noise must be introduced to mask the presence or absence of any input that possibly alters the output of an algorithm, even in the worst-case. This amount of noise can be unbounded, implying that an infinite amount of noise might be required to achieve this privacy definition. More generally, differential privacy is difficult to achieve for computations that are sensitive to the input of a single value over an unbounded domain, such as the reporting the mean, the smallest and largest order statistics, and other non-robust statistics that operate over the entire real line. Without additional modifications of the problem statement, such as placing domain restrictions or bounds on the input space or by replacing a sensitive algorithm with a more statistically-robust one, we cannot ensure that any arbitrary algorithm can be made differentially private with a finite (let alone reasonable) noise budget ϵ. While there are circumstances where such assumptions and modifications can be reasonably made in practice, we cannot immediately make use of differential privacy in every possible deployment scenario. It should be noted that this is not a bug in the differential privacy approach, but rather a feature, as it forces the algorithm designer to consider computational tools that are synergistic with the goal of not revealing "too much" about a single individual. It also suggests a need for thinking about privacy at the design stage and encourages the adoption of contextually relevant protection mechanisms.

4 The Policy Landscape

We briefly set forth provisions in a selection of applicable laws and policies which bear on how privacy requirements are set within practical data governance

regimes as they concern the design, engineering, and operation of database systems; because we are from the U.S. context, our summary is very U.S.-focused.

GDPR The European Union's new General Data Protection Regulation (GDPR) provides for a baseline standard of data protection and privacy that applies across all EU and European Economic Area (EEA) countries. The GDPR applies to all data about EU citizens, regardless of where in the world the data are collected or held, and violations carry maximum fines of €10 million or 4% of a company's global turnover. In addition to requiring traditional privacy notions based around confidentiality (e.g., collecting, storing, and processing only the data necessary to achieve some particular purpose and limiting processing only to claimed purposes), the GDPR provides a number of rights of interest to database design and operation. The general framework applies to "personal data", which is broadly speaking any data that can be tied to an individual person. While the law defines several classes of information which are affirmatively personal data, the question of whether functional identifiers should be treated as personal data remains unsettled [6]. Several special provisions apply to data processing defined as "profiling", which broadly covers any "automated processing of personal data" which "evaluate[s] certain personal aspects relating to a natural person".

Articles 9 and 10 restrict the processing of several classes of sensitive data for certain purposes, for example Article 9 stipulates that "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited" unless one of several conditions is met, such as the acquiring of affirmative consent. Article 10 prohibits processing based on criminal history except as allowed under the law of the relevant EU member state. But as noted above, the ease of inferring sensitive data makes it challenging to comply with these provisions, as covered personal data may be created or inferred during processing [37].

Articles 15–17 provide for *subject access, rectification, and erasure rights*, which enable subjects to demand all personal data about them, the ability to alter and correct those data, and the right to request that all such data be deleted. These rights apply whether data have been gathered from the subject directly or indirectly Implementing these processes efficiently represents a formidable engineering challenge in the era of large, distributed, federated databases. In particular, deletion may prove a challenge in systems where removing a single subject's data would prove costly or require denormalization or reindexing.

Article 20 provides for a right to data *portability*, which stipulates that subjects should be able to extract their data in open, machine-readable formats. This is more straightforward, but the nature of what formats are acceptable could benefit from input by the database community.

Articles 18, 21, and 22 provide for rights of the data subject to object to processing in a number of situations, including the use of data for marketing purposes and situations where processing is "solely automated". An active scholarly

and legal debate further asks whether these rights require that automated decisions be explained to data subjects and what that would require [33], although stronger explanatory rights may exist in Articles 13–15.

CCPA The California Consumer Privacy Act (CCPA) is a new state law in California, which takes effect on 1 January 2020 and provides a number of new consumer rights to California residents, including the right to access, rectify, and delete data similar to the GDPR. Additionally, the CCPA provides for data breach notification and the right to object to the sale of personal information, as well as the right to know when data are transferred to or shared with third parties and to opt out of most data sharing. Once again, the problematic issue of determining when data are personal data and thus covered under the law remains open to interpretation by both by technologists and in courts.

HIPAA The U.S. Health Insurance Portability and Accountability Act (HIPAA) governs the use of most medical and healthcare data in the United States. HIPAA provides a list of enumerated categories of "protected health information" (PHI) and a list of "covered entities" which must follow the rules, such as hospitals, insurance companies, and medical providers. The U.S. Department of Health and Human Services has created and administers the enforcement of a detailed HIPAA Security Rule, which defines exactly what data practices are and are not acceptable under the law. Broadly, however, HIPAA requires that patients give consent or receive notice for many uses of their data and that covered entities not share PHI without consent or to unauthorized other entities (covered or not).

The U.S. Federal Trade Commission Another tool in the policy landscape is are the unfairness and deceptive doctrines of Section 5 of the FTC Act. While not a privacy law per se, the unfairness and deception doctrines have become a major part of the US privacy policy landscape by policing behavior in the marketplace that is detrimental to consumers' privacy interests [15]. These flexible pieces of policy can be applied to many different situations, entities, and technologies. Under Section 5, an act is deceptive if there exists a material representation, omission, or practice that is likely to mislead consumers acting reasonably under the circumstances [11], whereas unfairness describes practices that cause or are likely to cause substantial injury to consumers that are not outweighed by the benefits to consumers or competition and are not reasonably avoidable [12].

5 Polystores as Privacy Opportunity: A Research Agenda

We have identified a gap between traditional approaches to privacy protection in databases and the modern understanding of privacy risks, especially those based on statistical inference. By facilitating cross-referencing of data across stores and faster deployment of models, polystore technologies have the potential to extend analytical capabilities and widen this gap. At the same time, we believe that the current interest in new polystore architectures presents an opportunity to create technologies that improve privacy protection in a meaningful way.

There is a natural alignment between the goal of protecting privacy and a polystore architecture. We now understand that privacy is not a property of

individual queries or even individual datastores. If we are to build any code that engages with privacy in a meaningful way, it will need the global visibility afforded by a polystore layer. As polystores enter operation, they will in turn provide a useful testbed for a range of privacy-preserving technologies

We identify three goals that make up a future-oriented research agenda for privacy in polystores. One is motivated by fair information practices and compliance with existing laws. The second and third are complementary responses to the advance of algorithms for inferring personal information.

5.1 Reconcile legal regimes with algorithmic capabilities

Our first research goal is motivated by the difficulty of mapping concepts from law onto the techniques of computer science. New privacy regulations like GDPR and the CCPA have created considerable uncertainty for companies, which have incurred significant expenditures in the hopes of being compliant without necessarily achieving confidence in their fidelity to legal requirements [18]. The meanings of many legal provisions are still being tested in courts, and may yet evolve in response to changing technologies. At the core of this issue are the very different languages used by the legal and computer science communities. Legal concepts like "singling out" are difficult to identify in technical architecture, and technical obstacles may conflict with the meaning anticipated in law [6].

Whether legal requirements can or should be formalized, and whether such formalizations are better conceptualized as necessary (as in Cohen and Nissim [6]) or sufficient (as in Nissim et al. [31]) conditions that approximate legal requirements is an inherently contextual question, based on the application scenario and the particular law at issue. Computer scientists often attempt to separate "policy" from "mechanism" in abstractions and subsequently assume that policy is given and the job of a mechanism is to ensure fidelity to that policy. However, such an approach is impoverished in the many important situations where acceptable outcomes and unacceptable outcomes are difficult to separate *ex ante* but must instead be established via *oversight and review* [20].

Laws are often specified in terms of flexible standards or general principles, which are applicable in many contexts and at different levels of abstraction, rather than actionable rules [13]. Because standards and principles are evaluated according to the balancing of countervailing concerns and may hinge on vague concepts such as standards of "reasonableness", they are generally not amenable to encoding as system requirements. Legal requirements are rarely, if ever, specific enough to be formalizable directly in code.

To address this gap, we foresee the need for interdisciplinary research teams, rooted in database technology and law. Such teams will be well positioned to understand the interface between computer systems and legal requirements, designing guarantees that support legal needs or compliance goals by providing relevant information or establishing key properties that can be consumed by the non-software processes of litigation, assessment, or other oversight rather than attempting to guarantee consistency with the law up front. For example, while it may be difficult to establish up front to what extent the use of sensitive classes

of data (e.g., gender, race, political affilations, or correllates of such attributes) constitutes illegal discrimination, systems can maintain query logs that aggregate estimates of how influential such categories have been based on models of the population in a database [1]. To match legal language to the capabilities of polystores will require explicit collaboration with legal scholars, who can interpret not only what laws are relevant and what these laws require, but how best to support their actual operationalization in real systems.

Interesting open questions in this area, prompted by the existing policy landscape include: Prompted by deletion rights in GDPR and CCPA and by the "right to be forgotten" in the GDPR, if a system can undo deletions, for example by restoring from a crash log or by losing deletion lists for write-only or write-mostly backing stores (e.g., tape libraries), when can data be safely considered deleted? If we delete an individual's data from a database, what about downstream uses of those data - can we track what computations or models derived from these data now require updating? Do machine learning models based on deleted data fall under these legal erasure rights? (The last of these is largely a legal question, but the database community must design in light of its uncertain answer.) Prompted by legal regimes that protect the use of sensitive data for significant computations (e.g., the GDPR) or which outlaw discrimination (several laws in many jurisdictions), must we carefully track the storage and use of data which can be used to make sensitive inferences, such as race, color, national origin, gender, religion, age, sexual activity and orientation, disability or health status, political affiliation, or genetic information? Additionally, prompted by subject access and correction rights in the GDPR, the CCPA, and sectoral laws in the US including HIPAA, FCRA, and ECOA, database systems should be prepared to answer queries that return all data about a particular individual and which allow these data to be corrected or deleted. Questions around schemata and workload support for such rights are important in practice.

5.2 Develop Accountability Mechanisms for Privacy Protection

Traditional privacy defenses, including access control, anonymization, as well as differentially private systems, can be viewed as preventative mechanisms. Like a fence or a moat, they are designed to foreclose the possibility of a privacy breach in advance, without the need for any active steps on the part of a system owner. Unfortunately, access control and anonymization are insufficient to counter modern attacks, and differentially private solutions may not be available for the vast majority of the computations that a typical company performs on data. The alternative to prevention, to borrow a term from the security literature, is accountability. An accountability system can be seen instead as an alarm: its purpose is not to prevent all privacy breaches in advance, but rather to make breaches detectable and attributable to individuals. Given the limitations of preventative systems, further research into accountability systems will be needed to provide privacy protection for many uses of databases in the future.

At its core, an accountability system for privacy needs the ability to measure the risk to individuals that arises from a given pattern of access. Such techniques

can be used to monitor the overall level of protection for users, to flag suspicious query behavior, and to support systems of deterrence for insiders. For sequences of queries that pose unusual privacy risks to individuals, such measures could trigger automatic suppression of results or assist firms in compliance efforts.

In certain cases, it may be appropriate to apply formal methods from differential privacy to measure risk in an accountability system. However, such methods are based on worst-case analysis and would lead to a conclusion of infinite privacy loss for many common database operations. To provide useful insight for these cases, new measurement techniques are required, which provide finite results in most cases, and are responsive to common human patterns of access that threaten privacy. Because such techniques will be based on a adversary model with limited capabilities, we will refer to them as *privacy heuristics*.

It is important to stress that privacy heuristics, by their nature, are not a perfect defense against the leaking of personal information. There is no limit to how clever a determined attacker may be in obfuscating an attempt to gain private information. Moreover, a higher standard of protection, including differential privacy, should be used for any output that is shared with the public. However, we believe that privacy heuristics have an important role to play in mitigating risks. It should be possible to recognize common patterns of behavior with high risks to individual privacy. Even in the case of a deliberate attack that aims to uncover an individual's secret, it may be possible to increase the cost to the attacker, or the probability of being discovered.

Future work on accountability systems for privacy can draw on diverse fields of study. New heuristics may be inspired by concepts from law, such as the GDPR's notion of "singling out." They may be informed by empirical studies of past privacy breaches, and responsive to particularly sensitive types of information. Search algorithms can be developed to detect when individual information can be extracted by differencing outputs across multiple queries or multiple islands. Finally, techniques from machine learning can be leveraged to recognize query patterns when an adversary deliberately obfuscates an attempt to gain private information. Taken together, we believe that these efforts can yield broad advances in how personal information is handled by firms and governments.

5.3 Incorporate Formal Privacy Techniques into Private Islands

As firms and governments seek to extract value from databases, they often encounter use cases for which informal protections based on accountability systems are insufficient. For some companies, it may be that sensitive data is accessed by too many employees to maintain a reasonable standard of accountability. At times, it may be necessary to present the results of an analysis to the public or to share data with outside collaborators.

To enable applications like these, we envision an effort to incorporate techniques from differential privacy into polystore systems. We note that the polystore layer is an appropriate place to deploy formally private algorithms, since it can maintain a view of what data is associated with individual units across various datastores. Algorithms can be naturally organized into formally private

islands that accept a limited range of commands, and inject randomness into all output before it is returned to the user. To accommodate the possibility of multiple private islands, a centralized privacy accountant is required to maintain a privacy budget and allocate it across islands, users, and queries. This accountant can be based on the ϵ of differential privacy, but also on alternative measures rooted in information theory or statistical inference.

The differential privacy literature contains a number of algorithms that can be immediately deployed in formally private islands. One strand of research concerns formally private algorithms that are appropriate for relational databases. This vein includes the PINQ [23] system, which provides analyst with a limited "SQL-like" language, as well as the system used by Uber for internal data analytics [17]. Private SQL has also been explored within federated database environments. Shrinkwrap is a private data federation that allows users of a system to have a differentially private view of other user's data that is a robust against computationally bounded adversaries [3]. When deploying formally private systems, care must be taken so that practical constraints of computing do not interfere with the theoretical privacy guarantees. For example, floating point representations can lead to privacy losses [24].

Another active area of differential privacy research concerns private implementations of machine learning algorithms. For algorithms that rely on stochastic gradient descent, a "bolt-on" implementation of differentially private gradient descent has been designed specifically to scale well for modern analytics systems [39]. For more general machine learning tasks, differentially private models can be constructed using the PATE framework [32]. Within federated systems, research is underway on the development of private federated learning methods that aim to construct machine learning models without specifically requiring individual user's data at runtime in a centralized location [4,35].

These existing solutions provide an excellent starting point for incorporating formal privacy standards and techniques into polystore systems. The intersection of these fields suggests a number of directions for possible future research. For one thing, work is needed to shed light on the proper design of a privacy accountant. The accountant bears the central task of allocating privacy budgets across individuals and across queries. Significant gains can be found by performing this role intelligently, directing privacy budget to important queries that require greater accuracy. In another direction, an advanced privacy accountant may recognize when queries are similar across different islands, and utilize shared randomness to save on privacy budget. Finally, work is needed on the design of interfaces that enable analysts to consider the privacy implications of their work and adjust their workflows to balance utility with privacy risk. This last direction would benefit from collaborations between database experts, differential privacy practitioners, and researchers in human-computer interaction.

References

1. Albarghouthi, A., D'Antoni, L., Drews, S., Nori, A.: Fairness as a program property. arXiv preprint arXiv:1610.06067 (2016)

2. Anderson, R.: Security engineering. John Wiley & Sons (2008)
3. Bater, J., He, X., Ehrich, W., Machanavajjhala, A., Rogers, J.: Shrinkwrap: efficient sql query processing in differentially private data federations. Proceedings of the VLDB Endowment **12**(3), 307–320 (2018)
4. Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., Rogers, R.: Protection against reconstruction and its applications in private federated learning. arXiv preprint arXiv:1812.00984 (2018)
5. Bruening, P.J., Waterman, K.K.: Data tagging for new information governance models. IEEE Security & Privacy **8**(5), 64–68 (2010)
6. Cohen, A., Nissim, K.: Towards formalizing the gdpr notion of singling out. arXiv preprint, arXiv:1904.06009 (2019)
7. Duggan, J., Elmore, A.J., Stonebraker, M., Balazinska, M., Howe, B., Kepner, J., Madden, S., Maier, D., Mattson, T., Zdonik, S.: The bigdawg polystore system. ACM Sigmod Record **44**(2), 11–16 (2015)
8. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. pp. 486–503. Springer (2006)
9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. pp. 265–284. Springer (2006)
10. Dwork, C., Rothblum, G.N., Vadhan, S.: Boosting and differential privacy. In: 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. pp. 51–60. IEEE (2010)
11. Federal Trade Commission: FTC Policy Statement on Deception. 103 F.T.C. 110, 174 (1984). `https://www.ftc.gov/public-statements/1983/10/ftc-policy-statement-deception`
12. Federal Trade Commission: FTC Policy Statement on Unfairness. 104 F.T.C. 949, 1070 (1984). `https://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness`
13. Feigenbaum, J., Weitzner, D.: On the incommensurability of laws and technical mechanisms: Or, what cryptography can't do. Security Protocols Workshop (2018)
14. Gellman, R.: Fair information practices: A basic history. Available at SSRN 2415020 (2017)
15. Hoofnagle, C.J.: Federal Trade Commission privacy law and policy. Cambridge University Press (2016)
16. Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B.C., Roth, A.: Differential privacy: An economic method for choosing epsilon. In: 2014 IEEE 27th Computer Security Foundations Symposium. pp. 398–410. IEEE (2014)
17. Johnson, N., Near, J.P., Song, D.: Towards practical differential privacy for sql queries. Proceedings of the VLDB Endowment **11**(5), 526–539 (2018)
18. Kamarinou, D., Millard, C., Oldani, I.: Compliance as a service. Queen Mary School of Law Legal Studies Research Paper, No. 287/2018 (2018)
19. Kohli, N., Laskowski, P.: Epsilon voting: Mechanism design for parameter selection in differential privacy. In: 2018 IEEE Symposium on Privacy-Aware Computing (PAC). pp. 19–30. IEEE (2018)
20. Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., Yu., H.: Accountable algorithms. University of Pennsylvania Law Review **165**(3) (2017)
21. Lee, J., Clifton, C.: How much is enough? choosing ε for differential privacy. In: International Conference on Information Security. pp. 325–340. Springer (2011)

22. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE'06). pp. 24–24. IEEE (2006)
23. McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. pp. 19–30. ACM (2009)
24. Mironov, I.: On significance of the least significant bits for differential privacy. In: Proceedings of the 2012 ACM conference on Computer and communications security. pp. 650–661. ACM (2012)
25. Nabar, S.U., Kenthapadi, K., Mishra, N., Motwani, R.: A survey of query auditing techniques for data privacy. In: Privacy-Preserving Data Mining, pp. 415–431. Springer (2008)
26. Naldi, M., D'Acquisto, G.: Differential privacy: an estimation theory-based method for choosing epsilon. arXiv preprint arXiv:1510.00917 (2015)
27. Narayanan, A., Felten, E.W.: No silver bullet: De-identification still doesn't work. Manuscript (2014)
28. Narayanan, A., Huey, J., Felten, E.W.: A precautionary approach to big data privacy. In: Data protection on the move, pp. 357–385. Springer (2016)
29. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large, sparse datasets. IEEE Security and Privacy (2008)
30. Narayanan, A., Shmatikov, V.: Myths and fallacies of personally identifiable information. Communications of the ACM **53**(6), 24–26 (2010)
31. Nissim, K., Bembenek, A., Wood, A.B., Bun, M.M., Gaboardi, M., Gasser, U., O'Brien, D., Vadhan, S.P.: Bridging the gap between computer science and legal approaches to privacy. Harvard Journal of Law and Technology **31**(2) (2018)
32. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755 (2016)
33. Selbst, A.D., Powles, J.: Meaningful information and the right to explanation. International Data Privacy Law **7**(4), 233–242 (2017)
34. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)
35. Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R.: A hybrid approach to privacy-preserving federated learning. arXiv preprint arXiv:1812.03224 (2018)
36. United States Department of Health, Education, and Welfare: Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens: Report. MIT Press (1973)
37. Wachter, S., Mittelstadt, B.: A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. Columbia Business Law Review (2018)
38. Warren, S., Brandeis, L.: The right to privacy. Harvard Law Review **4**(193) (1890)
39. Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., Naughton, J.: Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In: Proceedings of the 2017 ACM International Conference on Management of Data. pp. 1307–1322. ACM (2017)