Health and Technology (2019) 9:65–75
https://doi.org/10.1007/s12553-018-0250-6

ORIGINAL PAPER

Improving privacy preservation policy in the modern information age

John S. Davis II[1] · Osonde Osoba[1]

Received: 17 October 2017 / Accepted: 25 July 2018 / Published online: 21 August 2018
© IUPESM and Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract
Anonymization or de-identification techniques are methods for protecting the privacy of human subjects in sensitive data sets while preserving the utility of those data sets. In the case of health data, anonymization techniques may be used to remove or mask patient identities while allowing the health data content to be used by the medical and pharmaceutical research community. The efficacy of anonymization methods has come under repeated attacks and several researchers have shown that anonymized data can be re-identified to reveal the identity of the data subjects via approaches such as "linking." Nevertheless, even given these deficiencies, many government privacy policies depend on anonymization techniques as the primary approach to preserving privacy. In this report, we survey the anonymization landscape and consider the range of anonymization approaches that can be used to de-identify data containing personally identifiable information. We then review several notable government privacy policies that leverage anonymization. In particular, we review the European Union's General Data Protection Regulation (GDPR) and show that it takes a more goal-oriented approach to data privacy. It defines data privacy in terms of desired outcome (i.e., as a defense against risk of personal data disclosure), and is agnostic to the actual method of privacy preservation. And GDPR goes further to frame its privacy preservation regulations relative to the state of the art, the cost of implementation, the incurred risks, and the context of data processing. This has potential implications for the GDPR's robustness to future technological innovations – very much in contrast to privacy regulations that depend explicitly on more definite technical specifications.

Keywords Privacy · Digital privacy · Data privacy · Data utility · Anonymization · De-identification · Data management · HIPAA · GDPR

1 Introduction

The global data ecosystem underpins online commerce. That ecosystem involves individual persons[1] generating and providing personal data (the supply). Commercial entities and public institutions collect and process personal data (the demand) to provide some form of service. In many cases, the data collecting and data trading companies are 3rd party firms that are unknown to the individuals whose data is being collected; indeed, the individuals are often unaware that data is being collected.

Personalization and recommendation services offer concrete examples of this ecosystem in action. Users "permit" commercial platforms like Twitter, Netflix, and Amazon to store records of their behaviors in these spaces. In return, the platforms learn how to recommend or suggest items or services that users might like based on their past behavior data. The value to users is higher-level satisfaction. The value to platforms often comes in the form of some combination of higher user engagement, more attention to ads, more product sales etc.

But this data-for-service exchange process raises privacy concerns and forces society to balance the utility of the delivered services against the risk that the data may be used against the will of the individuals from which it is collected. On the one hand, the service offered provides communication, knowledge and commercial benefits that are widely valued.

[1] Or "natural persons", the terms that GDPR uses for individuals.

Regulation, G.D.P., 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, *59*, pp.1-88.

✉ John S. Davis, II
jdavis1@rand.org

Osonde Osoba
oosoba@rand.org

[1] RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, USA

On the other hand, privacy concerns are serious matters that are considered foundational in the laws of many governments.

The balance between the privacy of personal data and the utility the data affords is delicate, and arguably there is no domain in which this balance is more delicate than that of the healthcare industry. In healthcare, the data collected consists of patient and health provider records that are made available to pharmaceutical companies and the medical research community [1]. The service offered by the data collectors is the promised generation of medical insights with the potential of benefiting society at large. Regardless of the promises, the privacy risks involve sensitive health records that may literally reveal matters of life and death.

Given these privacy risks, many government agencies have mandated policies for constraining the exchange of personal data. In the United States, the exchange of healthcare data is subject to the Health Insurance Portability and Accountability (HIPAA). HIPAA is one of the most extensive privacy regulations in the U.S. and governs health data that is directly tied to an individual's identity by virtue of containing *personally identifiable information* (PII). If PII attributes have been removed from health data via de-identification or anonymization methods, HIPAA places no restrictions on exchanging the de-identified data. The main drawback of HIPAA is that extensive research has shown that anonymized data can be re-identified so that the identity of the original data subject is revealed.

Unfortunately, many privacy policies and data management regulations depend on anonymization as the primary mechanism for ensuring that sensitive data is not inappropriately released. There is a great need for privacy policymakers to treat anonymization, not as a monolithic technique, but to consider the range of methods that fit within an anonymization toolkit. More importantly, privacy policies must take into consideration the fact that anonymization provides no guarantees and often fails at maintaining privacy. What may have been effective several years ago is not well suited to the modern, data-based society in which we now live.

The purpose of this paper is to facilitate a nuanced discussion of anonymization for the privacy policy community to help better consider the options for balancing data privacy and data utility. In what follows, we present a discussion of anonymization as a mechanism for privacy preservation. First, we present the components of tabular data and how these components are modified through anonymization. We then review two classes of anonymization: syntactic anonymization and differential privacy. This is followed by a review of existing privacy policy that shows their connection to the various anonymization techniques. In particular, we review the European Union's General Data Protection Regulation (GDPR) and its alternative framing of privacy preservation. Recognizing that anonymization provides no privacy guarantees, we then present a consideration of alternatives to anonymization for achieving privacy.

2 The privacy of tabular data

Large quantities of human subject data are often represented as tables of rows and columns with one or more records (rows) per individual. Each entry contains a "tuple" of column values consisting of explicit or unique *identifiers* (e.g., Social Security numbers), *quasi-identifiers* (e.g., date of birth, race or zip code) and *sensitive attributes*[2] (e.g., salary or health conditions) [2]. One naïve approach to avoiding the release of private information is to remove unique identifiers from each entry. One problem with this approach is that makes an assumption on which columns should be considered sensitive. Consider the example tabular health data in Table 1. Presumably the health status column is sensitive; the DOB column might also be considered sensitive. Often sensitivity is tied to a notion of normalcy, a concept that varies based on population. The second problem with this approach is that the quasi-identifiers often provide enough information for an observer to infer the identity of a given individual via a "linking attack," in which entries from multiple, separate data sets are linked together based on quasi-identifiers that have been made public. Statistical disclosure researchers first expressed concerns about the re-identification risk posed by disclosing such quasi-identifiers [3]. Later Sweeney showed that 87% (216 million of 248 million) of the population in the United States could be uniquely identified if their 5-digit zip code, date of birth and gender is known [4]. Other work shows that, in general, subjects can be easily and uniquely re-identified using a very sparse subset of their data trails as recorded in commercial databases [5], especially if the data includes location and financial details. The result is that data subjects (the people the data represents) can be re-identified based on data that has had explicit identifiers removed.

Linking attacks are quite common and have led to the re-identification of data subjects in several high profile cases in which sensitive data was made public [6]. For example, AOL released web search query data in 2006 that was quickly re-identified by New York Times journalists [7]. Similarly, in 2006 Netflix released the de-identified movie ratings of 500,000 subscribers of Netflix with the goal of awarding a prize to the team of researchers who could develop the best algorithm for recommending movies to a user based on their movie history. Narayanan and Shmatikov showed that the Netflix data could be re-identified to reveal the identities of people associated with the data [8].

Given these high-profile cases, there has been active research into data anonymization techniques. Most data anonymization schemes attempt to preserve subject anonymity (non-identifiability) by obfuscating, aggregating, or

[2] The definition of what is sensitive depends on personal opinions and tastes, though many would agree that certain attributes would universally be considered sensitive.

Table 1 Example tabular data

Name	DOB	Country of Origin	Zip Code	Health Status
Jenn Jones	6/22/1954	USA	90003	Positive
Guy Walrand	8/14/1922	France	20622	Positive
Ali Mahmoud	11/22/1961	Canada	20610	Negative
Zhou Wei	3/17/1993	China	20610	Negative

suppressing sensitive components of the data set. The anonymization work has led to two related research areas: (i) privacy-preserving data publishing (also referred to as non-interactive anonymization systems) and (ii) privacy-preserving data mining (also referred to as interactive anonymization systems) [2, 9]. Non-interactive anonymization systems typically modify, obfuscate, or partially occlude the contents of a data set in controlled ways and then publish the entire data set. The publisher has no control of the data after publishing. Interactive anonymization systems are akin to statistical databases in which researchers pose queries to the database to mine for insights and the database owner has the option of returning an anonymized answer to the query. The AOL and Netflix cases are examples of non-interactive approaches and are the most common way to release date.

In addition to the dichotomy of privacy preserving data publishing and data mining, there are two general algorithmic approaches to anonymization: syntactic anonymization and differential privacy. Below we discuss both syntactic anonymization and differential privacy and show their relationship to privacy preserving data publishing and data mining. Later in this paper we will discuss additional models for privacy preservation, including alternatives to anonymization.

3 Syntactic anonymization

Syntactic anonymization techniques attempt to preserve privacy by modifying the quasi-identifiers of a dataset. The presumption is that if enough entries (rows) within a dataset are indistinguishable, the privacy concerns of the subjects will be preserved since each subject's data would be associated with a group as opposed to the individual in question. Manipulation of the quasi-identifiers can occur in a variety of ways, including via tuple suppression, tuple generalization and tuple permutation (swapping) so that a third party has difficulty distinguishing between separate entries of data.

The seminal approach to syntactic anonymization is the k-anonymity procedure [10]. It generalizes and suppresses components of data records such that any single disclosed record is indistinguishable from at least k other records. This effectively clusters the data into equivalence classes of minimum size k, making it difficult to resolve individual subjects better than these k-sized clusters. Further iterations on k-anonymity, such as l-diversity and t-closeness [11], attempt to buy more security by making the sensitive fields of the equivalence classes more statistically representative or more relatively uninformative to adversaries. Researchers spent the decade after its debut demonstrating that k-anonymity does not ensure privacy preservation; in response to such determination, several incrementally improved schemes were proposed: p-sensitive k-anonymity [12], l-diversity [13] and t-closeness [14].

Closer inspection shows that the goal of *k*-Anonymity and *l*-Diversity is to modify data releases to limit the amount of information an observer gains when starting from a state of background ignorance. These syntactic privacy approaches are susceptible to attack (e.g. linking and skewness) precisely because background ignorance varies depending on how much background knowledge exists. Consider as an example the variance in salary compared to the variance in political party affiliation in the US. Uninformed guesses on the latter are more likely to be accurate than uninformed guesses on the former (the accuracy of uninformed guesses being inversely related to background ignorance or entropy). *k*-Anonymity and *l*-Diversity aim to hedge against this sort of highly variable disclosure risk. That informs the choices of *k* and *l* in both approaches. So it is not surprising that they have not proven robust. *t*-closeness controls for a different sort of disclosure: it limits the amount of information observers can glean from comparing sensitive attribute distribution between full tables and their sub-groups. This is a more stable, achievable goal.

Each syntactic anonymization scheme has proven inadequate for privacy preservation [15] (although t-closeness comes close in theory, at the cost of limited data utility). These schemes all modify identifier and quasi-identifier fields to prevent observers from linking sensitive attributes back to unique users (re-identification). It has been argued that distinctions between identifiers, quasi-identifiers, and attributes (sensitive or otherwise) are at best artificial and potentially misleading; they present re-identification algorithms that are able to re-identify people using any type of distinguishing structured or unstructured signal, sensitive or otherwise [8]. They demonstrated these algorithms on Netflix movie ratings data and social network data.

3.1 *k*-Anonymity

A data table satisfies the *k-anonymity* property ([10, 14]) if every distinctly occurring sequence of quasi-identifiers has at

least k occurrences in the table. This means each record in the table is indistinguishable from at least k-1 other records with respect to the quasi-identifying fields.

Table 2 shows two sets of health records and is an example of k-Anonymization (for k=3) applied to the left data set and resulting in the right data set. The quasi-identifier tuple, (Age, Zip), uniquely identify records in the first table. The modifications to the quasi-identifier fields in the second data set ensure that all unique instances of the quasi-identifier tuple have at least 3 corresponding records. The modifications include syntactic actions like generalization (mapping the specific age "53" to the more general age range "[48-53]") and suppression (suppressing parts of the zip field). The table is now a collection of k-sized equivalence classes with respect to the tuple. This prevents observers from resolving past a group of k records using the quasi-identifier tuple as a key. Finding efficient and useful k-anonymizations is a computationally challenging task for k > 2 [16]. The Incognito algorithm was developed for partitioning tables to approximately satisfy k-anonymity [17].

k-Anonymous tables prevent identity disclosures but they do not prevent observers from learning attributes about individuals. For example, in Table 2, an observer can infer more precise information about a participant's relative risks for flu or cancer based on just background age data (a *background information attack*). Some k-anonymizations may result in equivalence classes with uniform distributions on the sensitive attribute. This leads to sensitive attribute disclosure for all records in those classes (a *homogeneity attack*).

3.2 *l*-Diversity

The l-diversity concept is an attempt to prevent homogeneity attacks [13, 14]. An equivalence class in a table satisfies the *l-diversity* property if the sensitive attribute has at least l well-represented values for the sensitive attributes in the record class. A table is l-diverse if every equivalence class is l-diverse. The concept of "l well-represented" sensitive values can have different meanings. For example, it could mean that there are l distinct values of the sensitive attribute (distinct l-diversity), or that the entropy of the sensitive attribute in each class is at least l bits (entropy l-diversity).

The k-anonymous table presented above satisfies 2-diversity (in the distinct l-diversity sense) on the sensitive attribute, "Diagnosis." Each k-sized equivalence class has at least 2 values for its sensitive diagnosis field: (measles, flu) or (cancer, obesity). Distinct l-diversity is identical to another k-anonymity variant, p-sensitive k-anonymity [12] when $l = p$. Finding l-diverse anonymizations is computationally difficult [11] and in practice harder than finding k-anonymizations for the same table. The Mondrian algorithm exists for partitioning tables to approximately satisfy l-diversity [18].

The quantity, l, is a measure of representativeness of the distribution of the sensitive attribute in the classes. It may not always be a raw count of the number of distinct values taken. Attribute entropy is another measure. The goal is to prevent leaking too much information about the relative frequencies on the sensitive attribute (like we did on the 3-anonymous table). But the group distributions of the sensitive property are often skewed enough compared to the overall table distribution. So observers are still able to make limited inferences about relative sensitive attribute propensities. This is a *skewness attack*, a generalization of the k-anonymity's homogeneity attack.

3.3 *t*-Closeness

The t-closeness approach to syntactic privacy aims to guard against a specific kind of information gain: information gained by comparing a t-close table release with the fully de-identified table (i.e. all quasi-identifier fields removed). A table satisfies t-closeness if its records are split into equivalence classes such that the distribution of sensitive attributes in the whole table and the equivalence classes of the t-close table are within t *distance units* of each other. This makes each equivalence class less distinguishable from the whole original table.

The distribution distance metric needs to be carefully chosen to be semantically sensitive. Li et al. [14] identify a metric that satisfies this constraint: the Earth Mover's Distance (EMD) metric. The EMD metric measures how much effort it takes to optimally convert the first probability distribution into the second. In the t-closeness case, how much effort it takes to transform the sub-group sensitive attribute distribution into the full table's sensitive attribute distribution. The table below (adapted from [14]) is 0.278-close on the Disease sensitive attribute. The Table 3 is also 3-diverse.

t-closeness only attempts to make sub-groups indistinguishable in sensitive distribution from the full table. This is a more stable, less-context-dependent goal. The intuition of controlling

Table 2 3-Anonymized health data

Age	Zip	Diagnosis
28	90145	Measles
21	90141	Flu
21	92238	Flu
55	92256	Cancer
53	92124	Obesity
48	92204	Obesity

Age	Zip	Diagnosis
[21-28]	9****	Measles
[21-28]	9****	Flu
[21-28]	9****	Flu
[48-55]	92***	Cancer
[48-55]	92***	Obesity
[48-55]	92***	Obesity

Table 3 *t*-Closeness anonymization for t=0.278

	Zip	Age	Disease
1	47677	29	gastric ulcer
2	47602	22	gastritis
3	47678	27	stomach cancer
4	47905	43	gastritis
5	47909	52	flu
6	47906	47	bronchitis
7	47605	30	bronchitis
8	47673	36	pneumonia
9	47607	32	stomach cancer

	Zip	Age	Disease
1	4767*	<40	gastric ulcer
3	4767*	<40	stomach cancer
8	4767*	<40	pneumonia
4	4790*	>40	gastritis
5	4790*	>40	flu
6	4790*	>40	bronchitis
2	4760*	<40	gastritis
7	4760*	<40	bronchitis
9	4760*	<40	stomach cancer

only the within-table information-gain means the t-closeness property is more robust for privacy-preservation. But it can severely reduce the utility of the released data. And, while checking for *t*-closeness is easy, enforcing t-closeness is computationally difficult [19]. Algorithms (like SABRE [20]) exist for creating tables that approximate the *t*-closeness property.

3.4 Differential privacy – a relative privacy promise

Differential privacy is motivated by the fact that the administrators of sensitive datasets have no control over the outside or background information available about dataset subjects. Differential privacy attempts to control the additional disclosure risk a participant incurs as a result of inclusion in the database, relative to available background information. The ideal differentially private database would reveal nothing about an individual that cannot be learned without access to the database. Differential privacy prepares for the case that if a subject's data will be added or removed from the database, the modification should not significantly change the overall statistics of the database. In effect, differential privacy does not strive for absolute secrecy but instead enables a candidate for inclusion in a differentially private database to rest assured that joining the database will not expose their data anymore than is currently the case prior to joining the database [21, 22].

Unlike syntactic anonymization, differential privacy modifies the actual values of the sensitive attributes (as opposed to the quasi-identifiers) by adding random noise with a judiciously chosen distribution. Ideally the noise will be such that privacy will be preserved without significantly changing the aggregate statistics of the dataset, which a legitimate inquirer may need to access.

In specifying differential privacy, there are three key concepts to consider. The first is the randomized function for adding noise to the dataset's sensitive attributes, referred to as a *mechanism*, M. M is simply a function that takes a tabular dataset as an input and produces a noisy, privatized result; this result should not differ too much from the input dataset so that analysis will still be accurate but should be noisy enough to prevent privacy breaches. Ideally if M is applied to two datasets that are close in value, the two results should have a high probability of being equivalent. This leads to the second key concept: a mechanism is described as being ε *differentially private* (or ε-dp) if the privatized results of applying M to two datasets that differ by only one row have probabilities that differ by a numeric value of ε.

The differential privacy of a randomized mechanism immediately begs the question of how much noise should be added to the data, what kind of noise and how does this relate to ε [23]? This leads to the third concept: the amount of noise depends on the *global sensitivity* of the query functions that will be applied to the data. For now, we'll consider only numeric query functions that operate on tabular datasets and produce a number, such as the min, max or average query. The global sensitivity of a query function is a measure of how much a single entry or row within a dataset will impact the value of the query function. The global sensitivity of a numeric query function is defined as the maximum difference between the outputs of the query function applied to all possible adjacent datasets; two datasets are *adjacent* if they are identical except for one row.

The global sensitivity places a constraint on how much noise a randomized mechanism can add to the contents of a dataset since at least enough noise must be added to obscure the maximum possible change of a query's output function. In general, differential privacy works best with query functions that have a small global sensitivity since this implies that privacy will be maintained without distorting the data too much. An example query function is the *count query* that returns the number of entries in a dataset that have a particular value. Some examples are how many entries represent subjects that have cancer or are college graduates. The global sensitivity of the count query is 1 since the addition of a single entry to a dataset will change the output of the count query by at most 1.

A challenge and criticism of differential privacy is that it has difficulty with query functions that have large global sensitivity. For example, the sensitivity of the average function (e.g., calculating the average salary of all entries) can potentially be unbounded. In effect, a differentially private system must have an a priori understanding of all possible values of sensitive attributes in order to calculate the amount of noise for the mechanism (as is discussed further in the sidebar). Another issue is that differential privacy can be computational untenable. From a privacy vs. utility standpoint, what a differentially private database gains in terms of privacy, it loses in terms of utility.

t-Closeness comes closest to differential privacy in motivation. t-Closeness only tries to minimize information gain relative to the whole table (as identified by the attribute distributions on the whole table). Differential privacy also only tries to minimize information gain relative to the whole table modified by single record deletions/modifications. Both are attempts to safeguard against relative disclosure instead of absolute disclosure risks. In contrast, k-anonymity and l-diversity do not take background knowledge about sensitive attributes into account; they try to prevent information gain relative to *any* background state of knowledge. This is generally infeasible since we cannot know what auxiliary information observers may bring to the data. Recent work [24] shows that t-closeness can be equivalent to ε-differential privacy in some data publishing contexts. Both approaches emphasize relative privacy over absolute privacy guarantees.

3.5 A differential privacy example

Consider a database that provides the average income of residents from a particular county. If you know that Mr. Gold Bags is preparing to move into the county, then querying the

database before and after his move would enable you to detect Mr. Bags' income. Differential privacy attempts to prevent this detection. We can think of differential privacy as enabling a form of plausible deniability for Mr. Bags – no one can prove that Mr. Bags' data is part of the database. As shown in Table 4, consider two datasets, D_1 and D_2, that are identical except that D_2 contains one row representing Mr. Bags' data. We can think of dataset D_1 as representing all entries of a database prior to the addition of Mr. Bags' data and D_2 representing all entries of a database after the addition of Mr. Bags' data. Since D_1 and D_2 differ by only one row (row 6 representing Mr. Bags' entry), we call them adjacent datasets.

In order for the database to be differentially private, we need to select a randomized function, a mechanism M, that adds noise to the datasets that will produce a randomized result R. Since D_1 and D_2 are adjacent, the probability that $M(D_1) = R$ should be close to the probability that $M(D_2) = R$. More formally we can write

$$P[M(D_1) = R]/P[M(D_2) = R] < e^{\varepsilon}$$

For small ε, note that $e^{\varepsilon} \sim 1 + \varepsilon$ and if our probabilities are identical, we get

$$1-\varepsilon < P[M(D_1) = R]/P[M(D_2) = R] < 1 + \varepsilon$$

The amount and kind of noise that M adds is constrained by the global sensitivity of the query function, f, that will be applied to the data. The global sensitivity can be written: $\Delta f = \max[f(D_1) - f(D_2)]$ for all possible adjacent datasets.

If we considered a count query, $\Delta f = 1$, since two adjacent datasets can different by at most 1. Dwork proved that noise with a Laplacian distribution (also called the symmetric exponential distribution) will maintain differential privacy if the value of Laplacian noise with parameter $b = \Delta f/\varepsilon$. Hence, a database in which the count query is applied will be differentially private if it uses a randomized mechanism that adds Laplacian noise with $b = 1/\varepsilon$.

The next question is what value of ε should we select. This choice is up to the differential privacy designers. The larger b is, the more noise we need to add in order to achieve differential privacy. Hence, a smaller ε provides more noise. As we increase Δf (greater global sensitivity), we need smaller ε to provide enough noise. Consider a query function that calculates the median salary. In this case, the global sensitivity is equal to the highest possible salary in the datasets (this is a worst case scenario). Multiple similar queries can also add up to reduce the privacy budget the added noise provides. This is a significant difficulty with differential privacy and has led to several other definitions of sensitivity, including local sensitivity and smooth sensitivity. The result is that while differential privacy is lauded for its ability to make privacy preservation guarantees, it has difficulty from a utility perspective. We discuss these difficulties more in Section (Syntactic vs Differential Privacy).

Table 4 Two datasets of financial data

Table D_1	
Row	Income
1	50,000
2	58,000
3	72,000
4	59,000
5	68,000

Table D_2	
Row	Income
1	50,000
2	58,000
3	72,000
4	59,000
5	68,000
6	350,000

4 Syntactic vs. differential privacy

In comparing the syntactic and differential privacy approaches, it is important to keep in mind the trade-off between privacy preservation and data utility. At the end of the day, datasets are shared to provide some utility (e.g., a research insight such as an understanding about the effectiveness of a medical procedure). At one extreme, all data can be released so that data utility will be maximized while privacy is completely violated. On the other extreme, not releasing any data will maximize privacy preservation but the shared data (an empty dataset) will be useless. Hence, an organization must seriously consider both their interests in data sharing and the risks that they are willing to accept.

An organization wishing to share sensitive data must consider the logistical challenges associated with the sharing process in addition to carefully balance the privacy and utility of released data. One key step in this consideration is the choice between syntactic and differential privacy. This choice also affects the choice between privacy-preserving data publishing (PPDP) versus privacy-preserving data mining (PPDM). As indicated above, in order to estimate the amount of anonymization noise (via a value for ε) differential privacy requires an understanding of the space of possible attribute values (even those not contained within the dataset in question) and the space of possible query functions that will be used to process the data (and that imply a value for global sensitivity) [25]. Hence, a case can be made that differential privacy is more amenable to privacy-preserving data mining in which the data administrator maintains control of the data and can limit the kinds of query functions that will be applied to the data. Syntactic anonymization techniques do not suffer from this constraint; with syntactic anonymization a dataset's quasi-identifiers are manipulated independently of the query functions or external data sources. Furthermore, if an organization chooses to support PPDM, they have the responsibility of hosting the interactive application through which the data mining queries are made available. There are exceptions to this reasoning, and several researchers have published work on how to use differential privacy techniques for non-interactive (PPDP) data releases [26, 27].

5 Anonymity in practice: Existing standards and legislation

The importance of privacy and anonymization for data sets is recognized in US and EU law. In the United States, much early work on the anonymization of data about individual respondents and statistical disclosure control was motivated by constitutionally-mandated privacy requirements for Census data-collection activity. There have been efforts to further enshrine privacy guarantees for sensitive data sets such as those needed for public health research or census efforts. These privacy regulations and protections tend to fall short of guaranteeing absolute privacy. This is, in part, a side effect of the research community's evolving understanding and facility with privacy-preservation measures. This non-absolutist frame also reflects the understanding evident in existing law that there needs to be a balance between an individual's right to privacy and the public utility that comes from having databases with personal information. Below are a few regulatory approaches to privacy that are applied by some of the most consequential government bodies (from a data quantity perspective), including examples from the United States and the European Union.

5.1 Health insurance portability and accountability act (HIPAA)

As stated above, HIPAA regulates, amongst many things, national standards for electronic health-data management. As enacted in 1996 by the United States Congress, its goals include a mandate to increase the efficiency of national health care and insurance systems while safeguarding the rights (including privacy) of patients. Title II of the Act is specifically concerned with health privacy. It places non-trivial constraints on the use of data sets containing PII, but it allows for the use of de-identified health information (DHI) without any regulatory constraints. The working definition of DHI is somewhat ambiguous in the act. The more descriptive of two definitions of DHI classifies data as DHI if it passes the "Safe Harbor" standard: a data set is considered DHI if it suppresses or generalizes references to all members of an enumerated set of 18 classes of identifiers, consisting of information such as names, social security numbers, and email addresses.[3]

Unfortunately, any coherent data trail, given enough ingenuity and effort, can be used to re-identify subjects. Explicit and non-evolving privacy regulations like HIPAA's Safe Harbor rule promotes privacy overconfidence while doing very little to protect subjects [28]. Attribute blacklists are thus inherently limited as privacy preservation tools. The current Safe Harbor rule incentivizes data publishers to meet Safe Harbor requirements and do no more than that. Publishers are thus inclined to treat the Safe Harbor rule as a suggestion instead of a test of minimal sufficiency. This is a predictable outcome given the tension between data privacy and data utility. The Safe Harbor rule effectively promotes a moral hazard: it lulls data publishers into a false sense of data security. An alternative approach to privacy regulation might compel data publishers to demonstrate that their privacy-preservation measures reasonably account for *state-of-the-art* re-identification techniques. GDPR attempts to address this gap.

[3] http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

5.2 Federal information security modernization act (FISMA)

The 2014 passage of FISMA extended the pre-existing mandate for privacy considerations by federal agencies of the U.S. government. Section IV of the act specifies the use of "privacy impact assessments" (PIAs) in which agencies conduct an evaluation of the privacy risks associated with their collection of personally identifiable information (PII). A given agency performs a PIA on its various initiatives and then must make the corresponding PIAs available to the public. The PIAs typically specify the kinds of data elements that will be stored about each individual, who will have access to the data elements and how such access may occur. For organizations that conduct PIAs and want to release their database for general use, the anonymization techniques mentioned above may be useful. The PIA report could indicate which data elements are considered unique identifiers, quasi-identifiers or simply sensitive along with an appropriate syntactic or differential privacy assessment of the dataset. However, the challenges described above in implementing these anonymization techniques may prevent their widespread adoption.

5.3 U.S. census bureau

The U.S. Census Bureau's mandate, powers, and restrictions are delineated in Title 13 of the U.S. Code, enacted in 1954. Title 13 USC §9 directs the Census Bureau to safeguard the privacy of the data they collect. More specifically, Title 13 USC §9(a.2) prohibits the Census Bureau from disclosing any data that can be used to identify individuals or establishments. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) reiterates this privacy-preservation requirement. This mandate compels the Census Bureau to avoid publishing any data that might be vulnerable to modern re-identification attacks. The Bureau has taken steps to satisfy this duty. It implements privacy impact assessments per the FISMA requirement mentioned above to ensure that any collected PII is both necessary and permitted.[4] The Bureau has

[4] The statement of this policy and links to archived PIAs can be found at: (http://www.census.gov/about/policies/privacy/pia.html). The PIAs also serve to record information-sharing partners (usually other federal agencies) and consent collection practices.

also done some research and published reports on internal statistical disclosure control practices [29]. These are mainly syntactic manipulation methods. For example, the Census data release system will withhold aggregate statistics if they represent a small enough set to pose disclosure risks.[5] This amounts to a version of syntactic anonymization (closest to *k*-anonymity). A recent Census data visualization webpage[6] reportedly [30] uses differential privacy preservation schemes.

Dwork and collaborators [31] argue that the statistical disclosure control framing is unsound and inadequate for modern privacy preservation. The application of syntactic anonymization for Census release applications may be sufficiently robust given the large numerical magnitude of typical released statistics. But, in principle, these kinds of releases are vulnerable to disclosures when the data is combined with secondary (possibly commercial) databases. Re-identification techniques are maturing quickly in the commercial sector. It is only a matter of time before some agent decides to bring these techniques to bear on Census data releases. This is especially true as cheap, high-powered computing proliferates. The Census Bureau may need to rethink its privacy safeguards to meet its Constitutional mandate. Rigorously speaking, the Bureau's mandate compels it to pay close attention to state-of-the-art re-identification methods.

5.4 The family educational rights and privacy act of 1974 (FERPA)

FERPA protects the privacy of student education records. FERPA is a federal law that applies to all schools that receive funds under an applicable program of the U.S. Department of Education. The law gives parents certain rights with respect to their children's education records, and these rights are transferred to the student when he or she reaches the age of 18 or attends a school beyond the high-school level. These rights include the right of parental consent for certain types of information disclosure, including the right to restrict the release of information associated with the student's education record.

FERPA allows non-consenting disclosure of records under three classes of conditions. The first class of non-consenting disclosure allows a school to disclose records to certain authorized parties such as school officials with legitimate education interest and accrediting organizations. The second class of non-consenting disclosure allows a school to disclose without consent "directory information," such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance, though such disclosure requires that the parents' be given notice and the right to request non-disclosure.

The third class of non-consenting disclosure allows a school to release education records from which personally identifiable information (PII) has been removed. The FERPA text defines PII as explicitly including the student's name; the names of the student's family members; personal identifiers such as social security numbers, student number or biometric records; indirect identifiers such as date of birth, place of birth and mother's maiden name; other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty; and information requested by a person whom the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates. The last two kinds of PII (i.e., linkable data and insider information) are sufficiently abstract as to raise questions about fair application. The removal of other PII in conjunction with non-consented release of directory information raises re-identification concerns. But FERPA's definition of PII could now be construed to limit any disclosure, given the current stated re-identification techniques and the availability of relevant linkable databases.

Several government bills have been issued that call for significant changes and/or supplements to FERPA[7]. These include the Student Privacy Protect Action (a FERPA rewrite), the SAFE KIDS Act and the Student Digital Privacy and Parental Rights Act of 2015 [32]. The bills contain a variety of provisions including new protections against exchange of student data and constraints on the ability for companies to use the data in advertising. As of this writing, none of these bills have passed.

6 The European Union's general data protection regulation (GDPR)

The European Union (EU) established the General Data Protection Regulation (GDPR) in April 2016, to govern the collection, storage, and use of data generated by "natural persons" (Article 1) under EU jurisdiction. GDPR is intended to create a uniform code of data practice across the EU commonwealth. The regulations recognized a fundamental right for the protection of data relating to natural persons (Article 2). It pronounces this right while maintaining that privacy rights are not absolute as they must be balanced against other fundamental rights and non-excludable public goods like security.

[5] The Census data releases on languages spoken at home and English-speaking ability demonstrate this approach (https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html). These tables withhold state-level statistics on low-use languages like Welsh or Papia Mentae.

[6] See, for example, the Census Bureau's online data visualization map: http://onthemap.ces.census.gov/

[7] http://www.nasbe.org/wp-content/uploads/2015-Federal-Education-Data-Privacy-Bills-Comparison-2015.07.22-Public.pdf

The GDPR is deliberately ambitious and forward-looking for a binding piece of regulation[8]. The intent, implementation, and feasibility of some of its stipulations have been subject to some debate already (e.g., a right to explanations[9] and affirmative consent procedures). More pertinent to privacy, it makes a distinction between *pseudonymized* and *anonymized* data (Article 4(5), Recitals 26-30). Pseudonymized data refers to the process we have called anonymization here: any data processing method that reduces the risk of re-identifying the records of individual natural persons. Anonymized data is defined to be the idealized setting in which data pertaining to natural persons has been stripped completely and is no longer identifiable. GDPR recognizes that pseudonymization is potentially reversible given enough resources (e.g. time, advances in computing power). But the *pseudonymization* and *anonymization* distinction seems unlikely to hold up with time.

GDPR is novel in the way it identifies the individual *natural person* as the primary holder of rights to most data related to the person. Data-processing entities accrue duties or responsibilities if they make use of user data. This reframes the discussion from what commercial or public entities *can* do with user data to what they are *allowed to do by users*. GDPR enshrines a number of rights for users:

1. Right of access and obtaining copies of data held by a processor or controller (Article #15)
2. Right to rectify errors in user data (Article #16)
3. Right to data erasure or "to be forgotten" (Article #17)
4. Right to restrict processing (Article #18)
5. Right to object to solely automated decision-making (Article #22 (3))
6. Non-binding right to explanations for decisions made by automated systems (recital #71)

Some scholars have argued that the right to explanations is non-binding since it is a recital not an article of the regulation. Others have argued that the combination of

– the right to contest automated decisions (in Article #22(3)) and
– the requirement for user access to "meaningful information about the logic involved" in an automated decision (Article #13(2f))

are sufficient basis for grounding the right to explanations discussed in recital #71. There are also very defined notice requirements in case of breaches (Articles #33/#34). These rights are not absolute. GDPR defines restrictions, for example for security (national & public) and criminal justice purposes (amongst a few others).

GDPR specifications are notably agnostic on the methodology of de-identification. It is a goal-oriented regulation; GDPR identifies privacy and other data rights as goals or ends-in-themselves and tries to define a regulatory infrastructure to safeguard these rights subject to considerations of the technological state-of-the-art, implementation costs, and the severity of potential disclosure risk. That regulatory infrastructure includes a Commission to oversee compliance and a technical advisory Board for making recommendations to the Commission on setting and updating technical details. GDPR also defines a set of blacklisted sensitive or protected data attributes (race, ethnicity, political opinions, beliefs, union membership, sexual orientation, genetic/biometric) similar to HIPAA's PII list. But these attributes are not central compared to the rights.

This agnostic approach is likely to enable the regulation's language to withstand future technologies that will impact the viability of anonymization (or lack thereof). Nevertheless, the Article 29 Working Party Opinion on anonymization techniques, a non-binding commentary associated with GDPR, did provide advisory recommendations for de-identifying data.[10] These include the addition of noise to personal identifiers and the replacement of personal identifiers with codes, as well as the application of *k-anonymity*, *l-diversity* and *differential privacy* techniques.

7 Alternatives to anonymity

There have been arguments for alternatives to anonymization that would avoid the computational challenges, the risk of third party re-identification and the impact on data utility. For example, researchers at MIT and Harvard drawing on their edX[11] experience showed that anonymization weakened the results of their analysis of student data. Accordingly, they suggested confidentiality policies that compel researchers with full data access to uphold the privacy of the human subjects [33].

There are several models that can inform alternatives to data anonymization for preserving privacy. One approach is to grant systematic researcher access on the premises of the data owning entity. For example, the U.S. Census Bureau has established 23 Research Data Centers[12] for accessing census data as well as data from over 50 other partner organizations.

[8] Zarsky, T.Z., 2016. Incompatible: The GDPR in the Age of Big Data. *Seton Hall L. Rev.*, *47*, p.995.
[9] Goodman, B. and Flaxman, S., 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.
[10] Article 29 Working Party, Opinion 05/2014 on Anonymization Techniques, WP216, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion- recommendation/files/2014/wp216_en.pdf
[11] edX is a provider of university-level massive open online courses, http://www.edx.org
[12] https://ask.census.gov/faq.php?id=5000&faqId=665

In order to access the RDCs, interested researchers must apply for Special Sworn Status[13] that allows access to unmodified census data under special regulations.

Another approach would treat sensitive research data similar to the way the U.S. federal government handles classified information [34]. This approach would require designated researchers to undergo a clearance process in order to gain access to sensitive data and could benefit from a well-established set of procedures for validating would-be researchers. Critics of the U.S. clearance process who argue that it is economically wasteful [35, 36] underscore the need to consider choices in up-front controls as well as penalties and enforcement in the case of violation.

The previous section on existing standards and legislation shows that a great deal of effort has been expended on protecting privacy through legislation. As the legislative examples illustrate, bright line specifications of PII risk falling prey to advances in data collection and state-of-the-art privacy research. Insights from other domains may offer useful instruction here such as security clearance classification systems[14]. In particular, the use of performance standards (indicating desired outcomes) as opposed to design standards (indicating the procedures used to meet a performance goal) may offer flexibility that will enable organizations to more easily achieve a desired privacy preservation goal.[15]

Another alternative is to not release data sets, but to set up a third-party entity that stores the data, and allows researchers to submit code – such as code to execute statistical algorithms or generate aggregate tables – to be run on the data, returning results once the absence of identifiable information has been verified [2]. Other approaches that deserve exploration include privacy preserving computational techniques such as fully homomorphic encryption and secure multiparty computation [37, 38]. While still nascent, these approaches may offer the potential for enabling analysis of sensitive data while not revealing the contents of the corresponding dataset.

8 Conclusion

The utility of data often stands in sharp contrast to the privacy of its subjects. The role of data has been transformed in every aspect of society so that data of all types holds indispensable value. At the same time, much data contains personally sensitive information that can result in reputational, financial and even physical harm if improperly used. A key assumption in our development as a data-based society is the assumption that we can preserve both privacy and utility. Furthermore, many government policies assume that although it is easier to privilege utility over privacy, there are de-identification techniques that will effectively preserve privacy when necessary.

This discussion puts that foundational assumption to the test. We provided an overview of recent anonymization approaches designed to protect our growing data stores, but the commensurate growth of powerful algorithms and hardware makes it increasingly easy for a motivated interloper to violate any reasonable assumption of privacy, even when attempts to anonymize data are employed.

This presents a problem for policy-makers. Key parts of government rely on the ability to protect data sets and the subjects in those data sets via anonymization. And many policy implementations address privacy preservation but arguably in only a perfunctory fashion.

The GDPR is a recent policy that takes a more sophisticated approach at addressing privacy preservation, by explicitly acknowledging the weaknesses associated with de-identification techniques and by avoiding regulatory language that presupposes particular technologies or de-identification techniques. The regulatory burden then moves over to the question of robust procedures for evaluating technological risk - in this case disclosure risk. GDPR tries to address this using a bicameral supervisory scheme (the Commission and the Board). The scheme's effectiveness for future-proof technology regulation will be interesting to observe.

Acknowledgements We would like to thank Marjory Blumenthal and Rebecca Balebako for their detailed and thoughtful review of early drafts of this document. We are immensely grateful for their comments and feedback. Any errors contained herein are our own and should not be attributed to them.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Tanner, Adam, Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records, Beacon Press. 2017.
2. G Cormode, D Srivastava. Anonymized Data: Generation, Models, Usage. *SIGMOD*, Providence, Rhode Island. 2009.
3. Dalenius T. Finding a needle in a haystack: identifying anonymous census records. J Off Stat. 1986;2(3):329–36.
4. L Sweeney. Uniqueness of Simple Demographics in the U.S. Population , LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA. Forthcoming book entitled, The Identifiability of Data. 2000.

[13] http://psurdc.psu.edu/content/applying-special-sworn-status
[14] https://en.wikipedia.org/wiki/Classified_information
[15] http://www.regblog.org/2012/05/08/the-performance-of-performance-standards/

5. de Montjoye Y-A, Radaelli L, Singh VK. Unique in the shopping mall: on the reidentifiability of credit card metadata. Science. 2015;347(6221):536–9.
6. Adam Tanner. Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study," *Forbes*, April 25, http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/print/. 2013.
7. Michael Barbaro and Tom Zeller. A Face Is Exposed for AOL Searcher No. 4417749, New York Times. 2006.
8. A. Narayanan, V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (SP '08). IEEE Computer Society, Washington, 2008, 111–125..
9. Clifton C, Tassa T. On syntactic anonymity and differential privacy. Trans Data Privacy. 2013;6(2):161–83.
10. Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertainty, Fuzziness Knowledge-Based Syst. 2002;10(05):557–70.
11. Dondi R, Mauri G, Zoppis I. The l-diversity problem: tractability and approximability. Theor Comput Sci. 2013;511:159–71.
12. Truta TM, Campan A, Meyer P. Generating microdata with p-sensitive k-anonymity property. Berlin: Springer; 2007. p. 124–41.
13. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. ACM Trans Knowled Discov Data (TKDD). 2007;1(1):3.
14. Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.
15. Domingo-Ferrer, Josep, and Vicenç Torra. "A critique of k-anonymity and some of its enhancements." *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*. IEEE, 2008.
16. Bonizzoni, P, Gianluca Della Vedova, and Riccardo Dondi. "The k-anonymity problem is hard." *Fundamentals of Computation Theory*. Springer Berlin Heidelberg, 2009.
17. LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Incognito: Efficient full-domain k-anonymity." *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005.
18. LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Mondrian multidimensional k-anonymity." *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006.
19. Liang, H, H Yuan. "On the complexity of t-closeness anonymization and related problems." In Database Systems for Advanced Applications, pp. 331-345. Springer Berlin Heidelberg, 2013.
20. Cao J, et al. SABRE: a sensitive attribute Bucketization and REdistribution framework for t-closeness. VLDB J. 2011;20(1):59–81.
21. Cynthia Dwork. Differential privacy: a survey of results. In Proceedings of the 5th international conference on Theory and applications of models of computation (TAMC'08), Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer-Verlag, Berlin, Heidelberg, 2008, 1-19.
22. Dwork C. An ad omnia approach to defining and achieving private data analysis. In: Bonchi F, Ferrari E, Malin B, Saygin Y, editors. Proceedings of the 1st ACM SIGKDD international conference on privacy, security, and trust in KDD (PinKDD'07). Berlin, Heidelberg: Springer-Verlag; 2007. p. 1–13.
23. Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science* (FOCS '07). IEEE Computer Society, Washington, DC, USA, 94-103.
24. Soria-Comas, Jordi, and Josep Domingo-Ferrer. Differential privacy via t-closeness in data publishing. *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on*. IEEE, 2013.
25. Sarathy R, Muralidhar K. Evaluating Laplace noise addition to satisfy differential privacy for numeric data. Trans Data Privacy. 2011;4(1):1–17.
26. Leoni, D. (2012), Non-interactive differential privacy: a survey., *in* Guillaume Raschia & Martin Theobald, ed., 'WOD' , ACM, , pp. 40-52.
27. G. Cormode, M. Procopiuc, D. Srivastava, and T. Tran. Differentially private publication of sparse data. In *International Conference on Database Theory (ICDT)*, 2012.
28. Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. 2010;57:1701.
29. Zayatz L. Disclosure avoidance practices and research at the US Census Bureau: an update. J Off Stat. 2007;23(2):253.
30. Klarreich, Erica. "Privacy by the Numbers: A New Approach to Safeguarding Data." *Quanta Magazine*. Quanta Magazine, 2012. Web. <https://www.quantamagazine.org/20121210-privacy-by-the-numbers-a-new-approach-to-safeguarding-data/>.
31. Chawla S, et al. Toward privacy in public databases. Berlin: Theory of Cryptography Springer; 2005. p. 363–85.
32. Roscorla, Tanya. "3 Student Data Privacy Bills That Congress Could Act On." *Center for Digital Education* March 24, 2016, http://www.centerdigitaled.com/k-12/3-Student-Data-Privacy-Bills-That-Congress-Could-Act-On.html
33. Daries JP, Reich J, Waldo J, Young EM, Whittinghill J, Seaton DT, et al. Privacy, anonymity, and big data in the social sciences. Queue. 2014;12(7):30. 12 pages
34. Access to Classified Information, Executive Order #12968, August 4, 1995, http://www.fas.org/sgp/clinton/eo12968.html
35. Dana Priest and William M. Arkin, "A hidden world, growing beyond control," Washington Post – Top Secret America, http://projects.washingtonpost.com/top-secret-america/
36. "White House orders review of 5 million security clearances," Nov 22, 2013, https://www.rt.com/usa/clapper-demands-security-clearance-review-173/
37. Gentry C, Halevi S. Implementing Gentry's fully-homomorphic encryption scheme. In: Paterson KG, editor. Proceedings of the 30th annual international conference on theory and applications of cryptographic techniques: advances in cryptology (EUROCRYPT'11). Berlin: Springer-Verlag; 2011. p. 129–48.
38. Lindell Y, Pinkas B. Privacy preserving data mining. In: Bellare M, editor. Proceedings of the 20th annual international cryptology conference on advances in cryptology (CRYPTO '00). London: Springer-Verlag; 2000. p. 36–54.