Working Paper

Privacy Preservation in the Age of Big Data

A Survey

John S. Davis II, Osonde A. Osoba

RAND Justice, Infrastructure, and Environment

WR-1161
September 2016

RAND working papers are intended to share researchers' latest findings and to solicit informal peer review. They have been approved for circulation by RAND Justice, Infrastructure, and Environment but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

For more information on this publication, visit www.rand.org/pubs/working_papers/WR1161.html

Published by the RAND Corporation, Santa Monica, Calif.

Privacy Preservation in the Age of Big Data: A Survey

John S. Davis II, Osonde Osoba[1]
RAND Corporation
1200 South Hayes Street
Arlington, VA 22202
{jdavis1, oosoba}@rand.org

Abstract

Anonymization or de-identification techniques are methods for protecting the privacy of subjects in sensitive data sets while preserving the utility of those data sets. The efficacy of these methods has come under repeated attacks as the ability to analyze large data sets becomes easier. Several researchers have shown that anonymized data can be re-identified to reveal the identity of the data subjects via approaches such as so-called "linking." In this report, we survey the anonymization landscape of approaches for addressing re-identification and we identify the challenges that still must be addressed to ensure the minimization of privacy violations. We also review several regulatory policies for disclosure of private data and tools to execute these policies.

Introduction

Data is the new gold. Enterprises rise and fall solely on their ability to gather, manage, and extract competitive value from their data stores. People are often the subject of these data sets—streaming video services collecting user movie interests, social media outfits observing users' commercial preferences, insurance agencies gathering information on users' risk profiles, or transportation companies storing users' travel patterns. Properly analyzed, data can yield insights that help enterprises, social services and the research community provide value. Such data often includes sensitive personal information. This is a key component of the data's value. For example, the U.S. Census Bureau compiles data about national demographics and socio-economic status (SES) that includes sensitive information (e.g., income). Similarly, data in other domains such as healthcare, education and social media contain sensitive information.

The data landscape sets up competing interests. The value of sensitive data sets is undeniable. Such data holds tremendous value for informing economic, health, and public policy research and a sizeable section of the tech economy is based on monetizing data. But the value of safeguarding user privacy during data release is equally undeniable. We have strong social, cultural, and legal imperatives requiring the preservation of user privacy. Organizations responsible for such data sets (e.g. in academia, government, or the private sector) need to balance these interests.

The key question becomes: Are there current techniques that can maintain the privacy of data subjects without severely compromising utility? We privilege privacy interests in this framing question because privacy often lags behind utility (sometimes only as an after-thought) in much of current data-analytic practice. And we focus on the data public release or publishing use-case. The rest of this discussion argues that the answer to our question is most likely "not really." Then we discuss some policy implications.

Brief Overview of Privacy Preservation

We can often visualize such data stores 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 or zip code) and s*ensitive*

[1] Authors listed alphabetically.

attributes[2] (e.g., salary or health conditions) [1]. One naïve approach to avoiding the release of private information is to remove unique identifiers from each entry. The 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 [13]. 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 [2]. 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 [7], 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. For example, AOL released web search query data in 2006 that was quickly re-identified by *New York Times* journalists [4]. 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 [6].

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 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) [1][8]. 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.

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 [9]. It generalizes and suppresses components of data records such that any single disclosed record is indistinguishable

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

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 [10], 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[42], l-diversity[41] & t-closeness[18].

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 subgroups. This is a more stable, achievable goal.

Each syntactic anonymization scheme has proven inadequate for privacy preservation [20] (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 [6]. They demonstrated these algorithms on Netflix movie ratings data and social network data.

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 [25][26].

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

k-Anonymity

A data table satisfies the *k-anonymity* property ([9][18]) 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.

Consider the following tables of health records:

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

The second table is an example of a 3-anonymization of the first (with respect to the quasi-identifiers, Age and Zip). The quasi-identifier tuple, (Age, Zip), uniquely identify records in the first table. The modifications to the quasi-identifier fields in the second table 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 ensures 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 [35]. The Incognito algorithm was developed for partitioning tables to approximately satisfy *k*-anonymity [38].

k-Anonymous tables prevent identity disclosures but they do not prevent observers from learning attributes about individuals. For example, in this table, 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*).

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 [28]? 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.

l-Diversity

The *l*-diversity concept is an attempt to prevent homogeneity attacks [41][18]. 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[42] when $l = p$. Finding *l*-diverse anonymizations is computationally difficult [34] and in practice harder than finding *k*-anonymizations for the same table. The Mondrian algorithm exists for partitioning tables to approximately satisfy *l*-diversity [37].

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.

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

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. [18] 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 [18]) is 0.278-close on the Disease sensitive attribute. The table is also 3-diverse.

	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	Brochitis
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

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 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 [33]. Algorithms (like SABRE [36]) exist for creating tables that approximate the *t*-closeness property.

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 [19] 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.

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) [24]. 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 [22][27].

Anonymization Tools

The implementation of privacy regulations and policies through any of the above anonymization techniques can be a complex endeavor and can lead to unexpected consequences that reveal sensitive data or impact data utility. To address these difficulties, data management tools have been developed to help an organization ascertain how the manipulation of identifiers will affect the efficacy of re-identification and calculate relevant parameters such as *k* or ε values. Well-designed tools could enable a better understanding of the impact of anonymization and could facilitate the communication of best practices. Unfortunately, many of the available anonymization tools target the research community and do not include turn-key offerings that are suitable for non-technologists or policy analysts.[3] Below we discuss two anonymization tools that are fairly accessible to the non-research community: ARX and PINQ. In addition, we review probabilistic databases that provide functionality that is similar to anonymization, though driven by a different motivation.

- *ARX* is a robust, open source tool for transforming tabular data to preserve privacy and runs as a standalone application on the three major operating systems [14]. It supports all of the major syntactic anonymization approaches, and it introduced support for differential privacy with ARX version 3.3. ARX includes a graphical user interface that allows a user to easily manage the columns of data by declaring them as either identifiers, quasi-identifiers, sensitive attributes or insensitive attributes. A particularly innovative aspect of ARX is its inclusion of features for assessing the utility of the dataset once anonymization has been applied. Hence, ARX enables a user to explore the trade-off as noise is added to the dataset. ARX is available on GitHub (https://github.com/arx-deidentifier/arx) via an Apache open source license.

- *PINQ* is an SQL-like implementation of a ε-differential privacy-preserving database query language published by Microsoft in 2009 [21]. It implements differentially private query mechanisms on regular data. It also provides a mechanism for managing the risks associated with calling a sequence of similar queries that can undermine ε-differential privacy guarantees. The source and an API are available.[4]

[3] An example of a research-oriented anonymization tool is the UTD Anonymization Toolbox [49].

[4] PINQ is available at http://research.microsoft.com/en-us/projects/pinq/

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. Consider two datasets (tables), 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.

Table D_1

Row	Income (U.S. Dollars)
1	50,000
2	58,000
3	72,000
4	59,000
5	68,000

⇨

Table D_2

Row	Income (U.S. Dollars)
1	50,000
2	58,000
3	72,000
4	59,000
5	68,000
6	350,000

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 $M(D_2) = R$ should be close to the probability that $M(D_2) = R$. More formally we can write

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

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

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

(Box continues on next page.) ⇨

Probabilistic Databases

While there are few tools designed explicitly for anonymizing data, there are several tools in the domain of probabilistic databases that are not intended to address privacy but provide features that are similar to anonymization tools. Probabilistic databases are used for a variety of applications, such as managing data from noisy sensor readings, outcomes from scientific experiments or data that is dirty due to manual entry errors. The noisiness resembles that of anonymized data. Several of the available tools for anonymizing data are reviewed below.

- *MayBMS* (pronounced "may b-m-s") is a probabilistic database developed by researchers at Cornell and Oxford University [15]. MayBMS is a complete database (as opposed to a dataset) and extends PostgreSQL (an open source database considered by many to be comparable to commercial databases). MayBMS does not provide facilities for adding noise to a dataset, but presumably it could be used to load anonymized data after noise has been added.

A Differential Privacy Example *(Continued)*

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 $b = \Delta f/\epsilon$. 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 $b = 1/\epsilon$.

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).

- *Orion* is a probabilistic database system developed at Purdue University [16]. A key feature of Orion is that it supports discrete uncertainty (e.g., uncertain membership in categories such as political parties) and continuous uncertainty (e.g., uncertain numeric values associated with a temperature sensor). Orion[5] runs on top of PostgreSQL.

- *MystiQ* is a prototype probabilistic database developed by researchers at the University of Washington. MystiQ is a database front-end that encodes imprecise information so that it can be stored in a standard relational database. MystiQ also translates non-probabilistic queries so that they can manipulate the probabilistic data stored in the underlying database. MystiQ is freely distributed and does not come with a license.[6]

- *BayesStore* is a probabilistic database system at UC Berkeley that was published in 2008 [17]. BayesStore introduces an underlying statistical model and applies a machine learning approaches to encode correlations that exist among the probabilistic data. An implementation of BayesStore has not been released as open source.

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

[5] Orion can be downloaded from http://orion.cs.purdue.edu/index.html

[6] MystiQ can be downloaded from http://homes.cs.washington.edu/~suciu/project-mystiq.html

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.

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.

U.S. Census Bureau

The U.S. Census' mandate, powers, and restrictions are delineated in Title 13 of the U.S. Code. 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 also implements privacy impact assessments per the FISMA requirement mentioned above to ensure that any collected PII is both necessary and permitted.[7] They also do some research and publish reports on internal statistical disclosure control practices[45]. These are mainly syntactic manipulation methods. For example, Census data release system will withhold aggregate statistics and their estimates if they are low enough to cause disclosure risks.[8] This amounts to a version of syntactic anonymization (closest to *k*-anonymity). A recent Census data visualization webpage[9] reportedly[44] uses differential privacy preservation schemes.

Dwork and collaborators [43] 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

[7] 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.

[8] 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.

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

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.

Health Insurance Portability and Accountability Act (HIPAA)

HIPAA regulates, amongst many things, national standards for electronic health-data management. 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 identifiers.[10]

Unfortunately, recent re-identification research (already discussed in preceding sections) shows that there is really no bright line between information (in general?) and personally identifiable information. 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 promote privacy overconfidence while doing very little to protect subjects [39]. 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 *bright line* test (or 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.

The European Union Data Protection Directive (EUDPD)

EUDPD, published in 1995, defines personal data as any information relating to an identified or identifiable natural person according to one or more factors specific to physical, physiological, mental, economic, cultural or social identity [40]. The Directive goes on to declare that processing of personal data shall mean any operation or set of operations performed upon personal data, including disclosure of personal information. This implies that the researcher and journalist disclosures cited above associated with AOL/New York Times and Netflix would be illegal under the Directive.

In May of 2016, the EU Directive became a Regulation and, hence, will be enforced as a law starting in 2017 in all EU countries (as opposed to the previous patchwork of slightly different laws across the EU countries) [50]. Ramifications of the 2016 update include the ability for users to make compensation claims, demand erasure rights (for having user data removed from datasets), and tighter rules on transferring data about EU users outside of the EU.

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

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

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.[11] 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 [46]. 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.

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[12] 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 [29].

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[13] for accessing census data as well as data from over 50 other partner organizations. In order to access the RDCs, interested researchers must apply for Special Sworn Status[14] 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 [30]. 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 [31][32] 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

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

[12] edX is a provider of university-level massive open online courses, http://www.edx.org

[13] https://ask.census.gov/faq.php?id=5000&faqId=665

[14] http://psurdc.psu.edu/content/applying-special-sworn-status

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.[15] 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.[16]

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 [11]. Other approaches that deserve exploration include privacy preserving computational techniques such as fully homomorphic encryption and secure multiparty computation [47][48]. While still nascent, these approaches may offer the potential for enabling analysis of sensitive data while not revealing the contents of the corresponding dataset.

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 people assume that although it is easier to privilege utility over privacy, with enough motivation, ingenuity, and forethought privacy can still be preserved. This discussion puts that foundational assumption to the test. We give an overview of recent privacy-preservation approaches designed to protect our growing data stores. But the commensurate growth of powerful algorithms and hardware makes it increasingly easy for a motivated analyst to violate any reasonable assumption of privacy.

This presents a problem for policy-makers. Key parts of government rely on the ability to protect data sets and the subjects in the data sets. And many policy implementations address privacy preservation but arguably in only a perfunctory fashion. Policy-makers need to start taking the current state of privacy-breaking technology into account to properly satisfy the privacy needs of citizens.

Acknowledgments

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.

References

[1] Graham Cormode and Divesh Srivastava, “Anonymized Data: Generation, Models, Usage,” *SIGMOD 2009*, June 29–July 2, 2009, Providence, Rhode Island.

[2] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000. Forthcoming book titled, *The Identifiability of Data*.

[3] Sweeney, Latanya. “*k*-anonymity: A model for protecting privacy.” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): pp. 557–570.

[4] Michael Barbaro and Tom Zeller, “A Face Is Exposed for AOL Searcher No. 4417749,” *New York Times*, August 9, 2006.

[5] Adam Tanner, “Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study,” *Forbes*, April 25, 2013, http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/print/

[15] https://en.wikipedia.org/wiki/Classified_information

[16] http://www.regblog.org/2012/05/08/the-performance-of-performance-standards/

[6] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP '08)*. IEEE Computer Society, Washington, DC, USA, pp. 111–125.

[7] de Montjoye, Yves-Alexandre, Laura Radaelli, and Vivek Kumar Singh. "Unique in the shopping mall: On the reidentifiability of credit card metadata." *Science* 347, no. 6221 (2015): pp. 536–539.

[8] Chris Clifton and Tamir Tassa. 2013. On Syntactic Anonymity and Differential Privacy. *Trans. Data Privacy* 6, 2 (August 2013), 161–183.

[9] Sweeney, Latanya. "*k*-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): pp. 557–570.

[10] Dondi, Riccardo, Giancarlo Mauri, and Italo Zoppis. "The *l*-diversity problem: Tractability and approximability." *Theoretical Computer Science* 511 (2013): 159–171.

[11] Graham Cormode and Divesh Srivastava, "Anonymized Data: Generation, Models, Usage," *SIGMOD 2009*, June 29–July 2, 2009, Providence, Rhode Island.

[12] Lambert, D., 1993. Measures of Disclosure Risk and Harm. *Journal of Official Statistics-Stockholm*, 9, pp. 313–313.

[13] Dalenius, T., 1986. Finding a needle in a haystack: Identifying Anonymous Census Records. *Journal of official statistics*, 2(3), pp. 329–336.

[14] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschlaeger, Klaus A. Kuhn, "ARX—A Comprehensive Tool for Anonymizing Biomedical Data," *Proceedings of the AMIA 2014 Annual Symposium*, November 2014, Washington D.C., USA.

[15] Jiewen Huang, Lyublena Antova, Christoph Koch, Dan Olteanu: "MayBMS: a probabilistic database management system." *SIGMOD 2009*: pp. 1071–1074

[16] Sarvjeet Singh, Chris Mayfield, Sagar Mittal, Sunil Prabhakar, Susanne Hambrusch, Rahul Shah. Orion 2.0: Native Support for Uncertain Data. *In Proc. of the ACM Special Interest Group on Management of Data (SIGMOD 2008)*, Vancouver, Canada, June 2008.

[17] Daisy Zhe Wang, Eirinaios Michelakis, Minos Garofalakis, and Joseph M. Hellerstein. 2008. BayesStore: managing large, uncertain data repositories with probabilistic graphical models. *Proc. VLDB Endow.* Volume 1, Issue 1 (August 2008), pp. 340–351.

[18] 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.

[19] 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.

[20] 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.

[21] McSherry, Frank D. "Privacy integrated queries: an extensible platform for privacy-preserving data analysis." *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data.* ACM, 2009.

[22] Leoni, D. (2012), Non-interactive differential privacy: a survey., in Guillaume Raschia & Martin Theobald, ed., 'WOD' , ACM, pp. 40–52.

[23] Rathindra Sarathy and Krish Muralidhar. 2010. Some additional insights on applying differential privacy for numeric data. In *Proceedings of the 2010 international conference on Privacy in statistical databases* (PSD '10), Josep Domingo-Ferrer and Emmanouil Magkos (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 210–219.

[24] Sarathy, R. and K. Muralidhar, "Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data," *Transactions on Data Privacy*, 4(1), pp. 1–17, 2011.

[25] Cynthia Dwork. 2008. 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, pp. 1–19.

[26] Cynthia Dwork. 2007. An ad omnia approach to defining and achieving private data analysis. In *Proceedings of the 1st ACM SIGKDD international conference on Privacy, security, and trust in KDD* (PinKDD'07), Francesco Bonchi, Elena Ferrari, Bradley Malin, and Yücel Saygin (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 1–13.

[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] 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, pp. 94–103.

[29] Jon P. Daries, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Daniel Thomas Seaton, Andrew Dean Ho, and Isaac Chuang. 2014. Privacy, Anonymity, and Big Data in the Social Sciences. *Queue* 12, 7, pages 30 (July 2014), 12 pages.

[30] Access to Classified Information, Executive Order #12968, August 4, 1995, http://www.fas.org/sgp/clinton/eo12968.html

[31] Dana Priest and William M. Arkin, "A hidden world, growing beyond control," *Washington Post–Top Secret America*, http://projects.washingtonpost.com/top-secret-america/

[32] "White House orders review of 5 million security clearances," Nov 22, 2013, https://www.rt.com/usa/clapper-demands-security-clearance-review-173/

[33] Liang, Hongyu, and Hao Yuan. "On the complexity of *t*-closeness anonymization and related problems." In *Database Systems for Advanced Applications*, pp. 331–345. Springer Berlin Heidelberg, 2013.

[34] Dondi, Riccardo, Giancarlo Mauri, and Italo Zoppis. "The *l*-diversity problem: Tractability and approximability." *Theoretical Computer Science* 511 (2013): 159–171.

[35] Bonizzoni, Paola, Gianluca Della Vedova, and Riccardo Dondi. "The *k*-anonymity problem is hard." *Fundamentals of Computation Theory*. Springer Berlin Heidelberg, 2009.

[36] Cao, Jianneng, et al. "SABRE: a Sensitive Attribute Bucketization and REdistribution framework for *t*-closeness." *The VLDB Journal* 20.1 (2011): 59–81.

[37] 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.

[38] 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.

[39] Ohm, Paul. "Broken promises of privacy: Responding to the surprising failure of anonymization." *UCLA Law Review* 57 (2010): 1701.

[40] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 31995L0046, http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046

[41] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M., 2007. *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), p. 3.

[42] Truta, T. M., Campan, A. and Meyer, P., 2007. *Generating microdata with* p-*sensitive* k-*anonymity property* (pp. 124–141). Springer Berlin Heidelberg.

[43] Chawla, Shuchi, et al. "Toward privacy in public databases." *Theory of Cryptography*. Springer Berlin Heidelberg, 2005. 363-385.

[44] Klarreich, Erica. "Privacy by the Numbers: A New Approach to Safeguarding Data." *Quanta Magazine*. 10 Dec. 2012. Web. <https://www.quantamagazine.org/20121210-privacy-by-the-numbers-a-new-approach-to-safeguarding-data/>.

[45] Zayatz, Laura. "Disclosure avoidance practices and research at the US Census Bureau: An update." *Journal of Official Statistics* 23.2 (2007): 253.

[46] 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

[47] Craig Gentry and Shai Halevi. 2011. Implementing Gentry's fully-homomorphic encryption scheme. In *Proceedings of the 30th Annual international conference on Theory and applications of cryptographic techniques: advances in cryptology* (EUROCRYPT '11), Kenneth G. Paterson (Ed.). Springer-Verlag, Berlin, Heidelberg, 129–148.

[48] Yehuda Lindell and Benny Pinkas. 2000. Privacy Preserving Data Mining. In *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology* (CRYPTO '00), Mihir Bellare (Ed.). Springer-Verlag, London, UK, UK, 36–54.

[49] UTD Anonymization Toolbox, University of Texas Dallas Security and Privacy Lab, http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php

[50] Hawthorn, Nigel, "10 things you need to know about the new EU data protection regulation," *Computer World UK*, May 6, 2015, http://www.computerworlduk.com/security/10-things-you-need-know-about-new-eu-data-protection-regulation-3610851/